REMARKS

The Official Action dated November 12, 2002 has been carefully considered. Accordingly, the changes presented herewith, taken with the following remarks, are believed sufficient to place the present application in condition for allowance. Reconsideration is respectfully requested.

By present amendment, in compliance with the Examiner's request, a paper copy of the "Sequence Listing" is entered into the specification. Also submitted on even date, to Box Sequence Listing, is a computer readable form (CRF) copy of the sequence listing. Applicants submit that the information recorded in the CRF is identical to the written sequence listing added by the present amendment. Additionally, it is believed that the neither the paper sequence listing nor the sequence listing in CRF include new matter, whereby entry is believed to be in order. Also by present amendment, references to particular SEQ ID NOs were added in compliance with 37 C.F.R. §1.821 to both the specification and claims 9 and 10. Claim 1 is amended in order to clarify that the molecule segment contributing to a disordered structure which is deleted is terminal, as previously recited in claim 4. In addition, the phrase "of the Class I Cytokine family" is added to claims 1 and 2, in order to define the invention. Support for this amendment can be found at page 6, lines 3-10, describing the scope of the inventive cytokines as comprising the hematopoietin receptor superfamily, which is commonly designated by those skilled in the art as the Class I Cytokine family. The preambles of claims 2, 6-10, 42 and 43 were amended to delete the word "modified" in order to clarify that the hGHR molecule itself is modified. It is believed that these changes do not involve any introduction of new matter, whereby entry is believed to be in order and is respectfully requested.

Claims 1, 2, 4-9 and 42-43 were rejected under 35 USC §112, first paragraph, as containing subject matter which was not described in the specification in such a way as to enable one skilled in the art to which it pertains to make and/or use the invention commensurate in scope with the claims. Specifically, the Examiner asserts that the instant specification, while being enabling for a modified human growth hormone receptor (hGHR) consisting of residues 32-237 or 32-234 of the native hGHR molecule, capable of being crystallized without being complexed to a ligand molecule, does not reasonably provide enablement for a cytokine receptor protein modified in the extracellular domain capable of

being crystallized without being complexed to a ligand molecule. More particularly, the Examiner asserts that the instant specification fails to provide any guidance as to how to generate crystals of a cytokine receptor which is modified in the extracellular domain by deletion of a molecular segment which contributes to a disordered structure. Further, the Examiner maintains that there is no evidence or sound scientific reasoning presented in either the instant specification or the prior art that would support a conclusion that such crystallization of any cytokine receptor is possible or was ever achieved because all the teachings of the instant specification are directed to a very specific segment of only one example of a cytokine receptor, which is human growth hormone receptor $hGHR_{1-237}$. The Examiner concludes that the instant specification fails to provide any guidance either on how to modify any given disclosed or, as yet, undiscovered cytokine receptor protein, or to permit an artisan to predict which segments of a receptor protein contribute to a disordered structure, thus necessitating an undue amount of experimentation in order to practice the invention.

This rejection is traversed. Applicants submit that the present amendments clarify the scope of the invention, and that the disclosure confers the ability to one of ordinary skill in the art to make and use the invention, commensurate with this scope. Therefore, the rejection is overcome and reconsideration is respectfully requested.

In particular, claim 1 is directed to a cytokine receptor protein of the Class I Cytokine family, modified in the extracellular domain, wherein at least one *terminal* segment which contributes to a disordered structure is deleted, the modified protein being capable of being crystallized without being complexed to a ligand molecule. Claim 42 is directed to a similar protein, specified as human growth hormone. Applicants submit that the determination of which segments of a molecule contribute to disordered structure is routine and predictable, thus providing precise guidance as to what constitutes the contemplated modifications that enable the invention. The Examiner's concern that the internal disordered regions critical to binding activity, which cannot be removed without negatively affecting binding activity, may not be determinable without undue experimentation, is moot. That is, the internal disordered regions that function to allow movement and binding conformation between rigid tertiary structures do not comprise the recited deleted segment, as claims 1 and 42 recite a terminal molecule segment which contributes to a disordered structure is deleted.

It is well-within the ability of one of ordinary skill in the art to determine which terminal regions of a Class I cytokine receptor contribute to a disordered structure.

Applicants submit herewith several pertinent publications which establish this ability, namely, (1) *"Structural Mechanisms for Domain Movements in Proteins"* Gerstein, Mark; Lesk, Arthur M.; Chothia, Cyrus, Biochemistry, Vol. 33, No. 22, pp. 6739-6749 (1994), (2) *"Improved Prediction of Protein Secondary Structure by use of Sequence Profiles and Neural Networks"* Rost, Burkhard; Sander, Chris, Proc. Natl. Acad. Sci. *USA,* Vol. 90, pp. 7558-7562 (1993), (3) *"Accuracy of Protein Flexibility Predictions"* Vihinen, Muano; Torkkila, Esa; Riikonen, Pentti, PROTEINS: Structure, Function, and Genetics, Vol. 19, pp. 141-149 (1994), (4) *"Hybrid System for Protein Secondary Structure Prediction"* Zhang, Xiru; Mesirov, Jill P.; Waltz, David, J. Mol. Biol., Vol. 225, pp. 1049-1063 (1992), (5) *"Rigid Domains in Proteins: An Algorithmic Approach to Their Identification"* Nichols, William; Rose, George D.; Ten-Eyck, Lynn; Zimm, Bruno H., PROTEINS: Structure, Function, and Genetics, Vol. 23, pp. 38-48 (1995), (6) *"Detection of Common Three-Dimensional Substructures in Proteins"* Vriend, Gerrit; Sander, Chris, PROTEINS: Structure, Function, and Genetics, Vol. 11, pp. 52-58 (1991), (7) *"Yeast Heat Shock Transcription Factor N-terminal Activation Domains are Unstructured as Probed by Heteronuclear NMR Spectroscopy"* Cho, Ho; Liu, Corey W.; Damberger, Fred F.; Pelton, Jeffrey G.; Nelson, Hillary; Wemmer, David E., Protein Science, Vol. 5, pp. 262-269 (1996), (8) *"Identifying Disordered Regions in Proteins from Amino Acid Sequences"* Romero, P; Obradovic, Z; Kissinger C.R.; Villafranca, J.E., Dunker, A.K., Proc.I.E.E.E. International Conference on Neural Networks, Vol. 1, pp. 90-95 (1997), (9) *"Protein Structure Prediction and Design"* Morea, Veronica; Leplae, Raphael; Tramontano, Anna, Biotechnology Annu Rev., Vol. 4, pp. 177-214 (1998), and (10) *"Predicting Protein Disorder for N-, C- and Internal Regions"* e-publication at http://www.jsbi.org/journal/GIW99/GIW99F04.pdf. There are repeated references within these articles to databanks of NMR and X-ray crystallographic-derived disordered regions within proteins. Drawing on this data, neural network models, which exploit the very predictability which the Examiner denies, have become ubiquitous. It is clear from inspection of these articles that determining, either via neural network models or empirically, whether a terminal region comprises a molecule segment which contributes to a disordered state is straightforward and can be done with a reasonable expectation of success.

Additionally, independent claim 1 is directed to Class I Cytokine Family receptor proteins. The striking homology of the extra-cellular domain of these proteins serves as the basis for this receptor classification, and, therefore, the basis inherently applies to those members of the family not yet discovered.

In summary, Applicants submit that a person of ordinary skill in the art has a reasonable expectation of success in practicing the present invention as defined by claims 1 and 42 because: 1) terminal segments of extracellular domains are precisely locatable regions, 2) deletion is easily accomplished by means well-known in the art, 3) regions contributing to disorder are readily identifiable, and 4) Class I Cytokines are defined by a structural conservation that confers a high degree of predictability as to the effect of structural modifications. Predictably, then, a person of ordinary skill in the art has a reasonable expectation that the receptor protein modified according to the present invention will be capable of being crystallized without being complexed to a ligand molecule.

It is therefore submitted that claims 1, 2, 5-9 and 42-43 are enabled by the specification in accordance with 35 U.S.C. §112, first paragraph, and that the rejection has been overcome. Reconsideration is respectfully requested.

Claims 1, 2 and 6 were rejected under 35 U.S.C. §112, second paragraph, as being indefinite. In particular, the Examiner asserts that claim 1 is indefinite because the recitation of "a molecule segment which contributes to a disordered structure" is considered vague and ambiguous, "since not every 'molecule segment which contributes to a disordered structure' is suitable for deletion to achieve crystallization."

This rejection is traversed. Applicants submit that this quote was intended to illustrate that, typically, internal disordered regions are relevant to binding conformation, and deletion may impact subsequent crystallization effort. However, as claims 1 and 42 recite a "terminal molecule segment which contributes to disordered structure", these claims are definite in accordance with 35 U.S.C. §112, second paragraph, and the rejection has been overcome. Reconsideration is therefore respectfully requested.

Claims 8-10 were rejected under 35 U.S.C.§112, second paragraph, as being indefinite. Specifically, the Examiner asserts that the preamble recitation "a modified growth hormone receptor" is vague and indefinite because it is not clear and cannot be determined from the claim what other possible modifications except truncation of the C-terminal end are encompassed by the claim.

This rejection is traversed. Applicants have amended claims 2, 5-10, 42 and 43 to remove "modified" from the preamble thereby clarifying the claims. Hence, the claims are

definite in accordance with 35 U.S.C. §112, second paragraph, whereby the rejection is overcome and reconsideration is respectfully requested.

The Examiner requested that Applicants submit a computer readable form (CRF) copy of a "Sequence Listing" which includes all of the sequences that are present in the instant application, a paper copy of the "Sequence Listing", an amendment directing entry of the paper copy into the specification, appropriate statements regarding content and the absence of new matter, and amendments to the instant specification and claims to comply with 37 C.F.R.§ 1.821(d) which requires reference to a particular SEQ ID NO in the specification and claims wherever a reference is made to that sequence. Without agreeing with the Examiner's basis, i.e., that the modifications of the claimed specific segment of hGHR are indefinite, Applicants believe they have fully complied with this request, as detailed *supra*.

It is believed that the above represents a complete response to the rejections and requests set forth in the Official Action, and places the present application in condition for allowance. Reconsideration and an early allowance are requested.

Respectfully submitted,

By: 

Holly D. Kozlowski, Reg. No. 30,468
DINSMORE & SHOHL LLP
Attorney for Applicants
1900 Chemed Center
255 East Fifth Street
Cincinnati, Ohio 45202
(513) 977-8568

875252v1

VERSION WITH MARKINGS SHOWING CHANGES MADE

In the Specification:

The specification is amended as follows:

The paragraph at page 2, line 19 through page 3, line 15 is amended as follows:

--According to a first aspect, the present invention is directed to a modified extracellular domain of a cytokine receptor protein, capable of being crystallized without being complexed to a ligand molecule. These modified proteins substantially maintain their activity to their native ligands and they will therefore constitute powerful tools for ligand interaction studies. The inventive, modified cytokine receptor preferably is of the type which oligomerizes when being bound to a ligand. This may include heterooligomerization of homodimerization, as discussed in *Mol. Cell. Biol.*, 1994, Vol. 14(6), p.3535-49: S Watowich et al. Most preferably, the modified receptor is a homodimeric cytokine receptor, such as the growth hormone receptor (hGHR) having an extracellular part consisting of 237 amino acids in its native state. The inventive proteins have at least one molecule segment contributing to a disordered structure deleted. Preferably, the deletion results in a truncation in at least one terminal end and most preferably it is truncated both in its C-terminal end and in its N-terminal end. More preferably, the inventive proteins are modified human growth hormone receptors (hGHR) with 31 or 33 amino acid residues removed in its N-terminal end and/or with 3 or 4 amino acid residues removed in its C-terminal end. Even more preferably, the inventive modified human growth hormone receptor (hGHR) consists of the amino acid residues 32-237 (SEQ ID NO: 2), 32-234 (SEQ ID NO: 3), or 34-233 (SEQ ID NO: 4) of the native molecule. Of these modified molecules, the truncated receptor consisting of amino acids 32-234 (SEQ ID NO: 3) of the native molecule is the most preferred. It should be emphasized that said modified cytokine receptors would be readily produced by the skilled person with existing methods of recombinant technology and their production in a recombinant host and their subsequent purification, therefore are not parts of the present invention. Further aspects of the invention are disclosed below.--

The paragraph at page 12, line 8 through page 14, line 2 is amended as follows:

--The hGH and hGHR used in the protein crystallographic work were expressed and purified as previously described in Sundström et al, (1996). Truncation mutants of hGHR

were created using standard sub-cloning techniques and the expressed protein was assayed for hGH binding using affinity and size exclusion gel filtration chromatography as well as BIAcore (Pharmacia Biosensor, Sweden) measurements. The $hGHR_{32-234}$ (SEQ ID NO: 3) protein was crystallized by vapor diffusion, using 3 ml protein solution (7mg/ml in 10 mM ammonium acetate) mixed with 3 ml of 0.33 M NH_4SO_4 30% (w/v) PEG-2000-dimethyl ether, 1% (v/v) DMSO and 100 mM MES buffer at pH 6.4 in a sealed tissue culture 24-well plate (Falcon, USA). The crystallization droplets were equilibrated at +18°C with 1 ml of the mother liquor for 2-4 weeks to obtain optimal quality crystals that diffracted to at least 2.9 Å with a conventional X-ray source. The crystals were frozen directly in the N_2 beam by adding a 1:1 mixture of 25% (v/v) ethylene glycol and 25% glycerol (v/v) to the crystallization droplet. Data was collected at station A1 at Cornell High Energy Synchrotron Source using a CCD detector (Area Detector Systems Corp., USA). The data was indexed, processed and scaled in the tetragonal spacegroup I4 using the programs DENZO and SCALEPACK, developed by S. Bailey in the SERC Daresbury Laboratory, Warrington, 1993. A molecular replacement search procedure was performed using the program AMORE, also developed by Bailey, 1993. The co-ordinates of the site 1 binding hGHbp molecule in our 2.5 Å hGH:hGHR 1:2 complex was used. The highest scoring solution in the resolution interval 8 - 4 Å was found in space group $I4_1$ with two hGHbp molecules in the asymmetric unit. A rigid-body refinement in X-plor, described by J. Navaza in *Acta Cryst.*, 1994, Vol. A (50), pp. 157-163, with individual hGHbp domains including data between 10-6, 10-5 and 10-3.5 Å in each respective cycle, decreased both the R- and Free-R values (described by A.T. Brünger in *Nature*, 1992, Vol. 355, pp. 472-475) dramatically when compared to previous runs where the native hGHbp domain arrangement was used. A cyclic process of model building in O, described by Brünger, 1992, followed by NCS restrained POWELL minimization in X-plor, using data between 15 - 2.3 Å, which was corrected for most main and side chain changes to the search molecule. At this stage, the first simulated annealing run, as described by T.A. Jones, et al. in *Acta Cryst*, 1991, Vol. A(47), pp. 110-119, was performed using a slow-cooling protocol from 3000 K to 300K in 50 Ps steps. Solvent molecules were introduced into FoFc densities above 3.0 s. After 3 cycles, a total of 327 solvent molecules had been introduced and assigned to the protein chain using the programs DISTANG and WATERTIDY developed by A.T. Brünger et al., 1989, in the CCP4 program package. A final POWELL minimization was performed, followed by a simulated annealing run from 2500 K to 300 K in 50 ps steps and including data between 15 to 2.3 Å. Individual B-value refinement was added as the final step, and solvent molecules with high

temperature factors, greater than 50 $Å^2$, or absent 2FoFc electron densities cut-off at 1.0 s, were removed. The Free-R value was used to validate the progress of the entire refinement. The final model consisted of residues 32 - 52, 63 - 70 and 80 - 234 of both molecules in the asymmetric unit as well as 261 solvent molecules and two sulphate ions. At the present stage of refinement, the R-factor of the model is 21.7% (R-free 29.3%), using data between 10 - 2.3 Å. As a control, a dataset to 3.2 Å at room temperature was collected. No significant differences to the 2.3 Å structure were observed, showing that the transfer to cryogenic conditions did not induce conformational adaptation. See also, Merritt et al, *Acta. Cryst.*, D50, 869-73 (1994).--

The Table 1 heading at page 15, line 3 is amended as follows:

--Crystallographic data for $hGHR_{32-234}$ (SEQ ID NO: 3)--.

In the Claims:

Claims 1, 2, 5-10, 42 and 43 are amended as follows:

1. (Twice amended) A cytokine receptor protein of the Class I Cytokine family, modified in the extracellular domain, wherein at least one terminal molecule segment which contributes to a disordered structure is deleted, the modified protein being capable of being crystallized without being complexed to a ligand molecule.

2. (Amended) A [modified] protein according to claim 1 being a homo- or heterodimeric cytokine receptor of the Class I Cytokine family.

5. (Amended) A [modified] protein according to claim [4] 1 truncated in its C-terminal and in its N-terminal end.

6. (Twice Amended) A [modified] protein according to claim 5 wherein the cytokine receptor protein is human growth hormone receptor (hGHR).

7. (Twice Amended) A [modified] human growth hormone receptor protein (hGHR) according to claim 6 having 31 or [32] 33 terminal amino acid residues removed in its N-terminal end.

8. (Third amendment) A [modified] human growth hormone receptor protein (hGHR) according to claim 6 having 3 or 4 terminal amino acid residues removed in its C-terminal end.

9. (Fourth amendment) A [modified] human growth hormone receptor (hGHR) consisting of residues 32-237 (SEQ ID NO: 2), 32-234 (SEQ ID NO: 3), or 34-233 (SEQ ID NO: 4), of the native hGHR molecule.

10. (Third amendment) A [modified] human growth hormone receptor (hGHR) according to claim 9 consisting of residues 32-237 (SEQ ID NO: 2), of the native hGHR molecule.

42. (Amended) [Modified human] Human growth hormone receptor protein, comprising human growth hormone receptor protein truncated in at least one terminal end to delete at least one molecule segment which contributes to a disordered structure, the modified human growth hormone receptor protein being capable of being crystallized without being complexed to a ligand molecule.

43. (Amended) [A modified human] Human growth hormone receptor protein according to claim 42, truncated in its C-terminal end and in its N-terminal end.

Biochemistry

Volume 33, Number 22 June 7, 1994

Perspectives in Biochemistry

Structural Mechanisms for Domain Movements in Proteins†

Mark Gerstein,*,‡,§ Arthur M. Lesk,*,‖ and Cyrus Chothia*,‡,⊥

MRC Laboratory of Molecular Biology, Department of Haematology, Cambridge University, and Cambridge Center for Protein Engineering, Hills Road, Cambridge CB2 2QH, U.K.

Received February 15, 1994; Revised Manuscript Received April 1, 1994®

ABSTRACT: We survey all the known instances of domain movements in proteins for which there is crystallographic evidence for the movement. We explain these domain movements in terms of the repertoire of low-energy conformation changes that are known to occur in proteins. We first describe the basic elements of this repertoire, hinge and shear motions, and then show how the elements of the repertoire can be combined to produce domain movements. We emphasize that the elements used in particular proteins are determined mainly by the structure of the interfaces between the domains.

Nearly all large proteins are built from domains (Wodak & Janin, 1981), and large relative movements of domains provide spectacular examples of protein flexibility. Domain motions are important for a variety of protein functions, including catalysis, regulation of activity, transport of metabolites, formation of protein assemblies, and cellular locomotion. Domains often close around a binding site between them. Generally, the presence of bound substrates stabilizes a closed conformation, and their absence favors an open conformation. Consequently, domain motions illustrate induced fit in protein recognition (Koshland, 1958).

Most of our information on the mechanisms of domain movements has come from X-ray crystal structures of open and closed conformations of particular proteins. The results of early investigations were reviewed by Janin and Wodak (1983) and by Bennett and Huber (1984). Since then, a considerable amount of new information has become available, and we review here the portion of this information that concerns structural mechanisms of domain closure.

In catalysis, domain closure often excludes water from the active site and helps position catalytic groups around the substrate. It also traps substrates and prevents the escape of reaction intermediates (Anderson *et al.*, 1979; Knowles, 1991). Domain closure, therefore, must be fast, and the transition between open and closed forms cannot involve high-energy barriers. Protein interiors, however, have features that place strong constraints on their possible conformational changes: they are close-packed with main chains and side chains in preferred conformations and with buried polar groups hydrogen bonded. In the first part of this review, we discuss the repertoire of possible low-energy conformational changes that are available to proteins, *i.e.*, their intrinsic flexibility. In the second part we describe how this repertoire of low-energy conformational changes are used to produce domain movements in particular proteins.

THE INTRINSIC FLEXIBILITY OF PROTEINS

The intrinsic flexibility of proteins is taken here to mean the ability of different *segments* of the protein to move in relation to one another with only small expenditures of energy. Analysis of protein crystal structures has shown that this intrinsic flexibility can take two forms: hinge motions in strands, β-sheets, and α-helices that are not constrained by tertiary packing interactions and shear motions between close-packed segments of polypeptide (Figure 1; Table 1).

(A) Hinge Motions in Strands, Sheets, and Helices Not Constrained by Packing Interactions. (1) β-Strands. The most basic motion of a polypeptide chain is a few large changes in main-chain torsion angles in a localized region, *i.e.*, at a

† Supported by Damon Runyon–Walter Winchell Fellowship DRG-1272 (M.G.) and the Kay Kendall Foundation (A.M.L.).

‡ MRC Laboratory of Molecular Biology.

§ Present address: Beckman Center for Structural Biology, Department of Cell Biology, Stanford Medical School, Stanford, CA 94305.

‖ Department of Haematology, Cambridge University.

⊥ Cambridge Center for Protein Engineering.

® Abstract published in *Advance ACS Abstracts*, May 1, 1994.

0006-2960/94/0433-6739$04.50/0

Table 1: Comparison of Hinged vs Shear Mechanisms for Domain Closure

	shear mechanism	hinged mechanism
simple example	citrate synthase	lactoferrin
main-chain packing	constrained by close packing	free to kink
main-chain torsions	many small changes	a few large changes
motion overall	concatenation of small local motions	identical to twisting at hinge
motion at interface	parallel to plane of interface (shear)	perpendicular to plane of interface, exposing and burying surfaces
side-chain packing	same packing in both forms	new contacts created; packing at base of hinge crucial
side-chain torsions	mostly small changes	some large changes

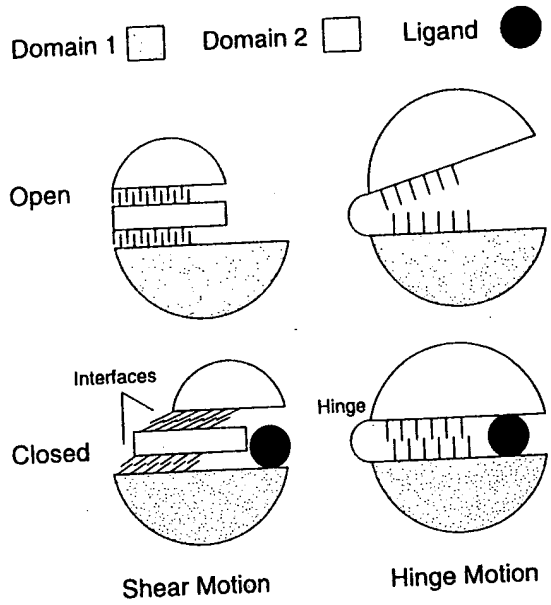

FIGURE 1: Hinged and shear mechanisms for domain closure. See Table 1 for a summary of the characteristics of both mechanisms.

hinge. The deformation of an extended strand is the simplest hinge motion because its only constraint is that the torsion angles of the strand remain in the allowed regions of the Ramachandran diagram. Consequently, its torsion angle changes can be very large and the resulting motion can rotate the polypeptide chain up to 60°. As shown in Figure 2A, in lactate dehydrogenase two adjacent torsion angle changes rotate a strand by ~35° in a direction not accessible by a single change (Gerstein & Chothia, 1991).

(2) β-Sheets. Two strands connected in a β-sheet can move together like the hinges on a door. However, the necessity that the strands remain hydrogen bonded together provides an additional structural constraint beyond the limitations of the Ramachandran diagram. As shown in Figure 2B, for the hinged sheet in lactoferrin this additional constraint means that in both strands the rotation axes of the principal torsion angle changes must be nearly parallel to each other and to the axis of the overall rotation of the sheet (Gerstein *et al.*, 1993b). Three large (>30°) torsion angle changes produce the bulk of the motion, rotating the sheet by 53°.

(3) α-Helices. Hinges in α-helices present a contrasting story. Because residues in helices are subject to more severe hydrogen-bonding and steric constraints than those in sheets, their torsion angles are restricted to a smaller region of the Ramachandran diagram. Thus, if residues are to remain in a helical conformation, the possible changes in their torsion angles are correspondingly smaller than those of residues in an extended conformation, and the deformation of helices must be spread over more residues than the deformation of sheets. Such spread-out helical deformations can produce bending motions: eight torsion angle changes between 9° and 15° in the C-terminus of a helix in a mutant lysozyme bend its end to produce a shift of 3.3 Å (Dixon *et al.*, 1992; Figure 2C). Similar deformations can also stretch a helix: six torsion angle changes over four residues at the N-terminus of a helix in lactate dehydrogenase tighten the helix up from an α to a 3_{10} conformation and stretch it by 3.3 Å (Gerstein & Chothia, 1991).

A different situation occurs in those helices that contain kinks, which often involve prolines. The disruption in the normal pattern of hydrogen bonding, and hence in the constraints on the helix, allows larger torsion angle changes. As shown in Figure 2D, such large torsion angle changes have been found in the proline-kinked helix in adenylate kinase.

The interconversion of helical and extended conformations is also possible and has been found in calmodulin (Ikura *et al.*, 1992; Meador *et al.*, 1992, 1993) and triglyceride lipase (Derewenda *et al.*, 1992). While such an interconversion may involve crossing energy barriers somewhat higher than those in the motions discussed above, it permits large torsion angle changes and large deformations. In calmodulin, torsion angle changes in five residues in the middle of a long helix split it into two smaller helices, separated by four residues of extended strand. These two small helices are inclined at an angle of ~100°.

(B) Limited Shear Motions of Close-Packed Segments of Polypeptide. The preceding discussion of hinges considered only the effects of structural constraints intrinsic to β-strands, β-sheets, and α-helices—*i.e.*, constraints arising from the requirements of secondary structure. The interactions that stem from tertiary structure provide even more severe structural constraints. Most of the atoms in a protein are partially buried and closely packed—in particular, most of the main chain is buried beneath layers of side chains. This close packing precludes large torsion angle changes and hence hinges. Indeed, a structural requirement for a residue to act as a hinge is that it have few tertiary structure packing constraints on its main chain.

As shown in Figure 3, we can divide movements of close-packed segments of polypeptide into those that are perpendicular to an interface and those that are parallel. Hinges from outside the region of the interface can produce a motion perpendicular to the plane of an interface (so the interface exists in one conformation but not in the other, as in the opening and closing of a book). As discussed below, this sort of motion can be driven by ligands stabilizing a closed conformation. Motions parallel to the plane of the interface are limited by the packing contacts involving the interdigitation of side chains. Large shifts of close-packed segments of polypeptide would require switching between different interdigitating configurations. Although such packing changes are seen at the subunit interfaces of allosteric proteins (Perutz, 1989), they have not been observed, so far, in domain closure. This is probably because such motions involve crossing high-energy barriers and would not occur with sufficient rapidity.

Small shear motions (Figure 3) that do not involve repacking of the interface are commonly involved in domain closure and have the following characteristics: (1) Interdigitating side chains accommodate shear motions, mostly, by small (<15°)

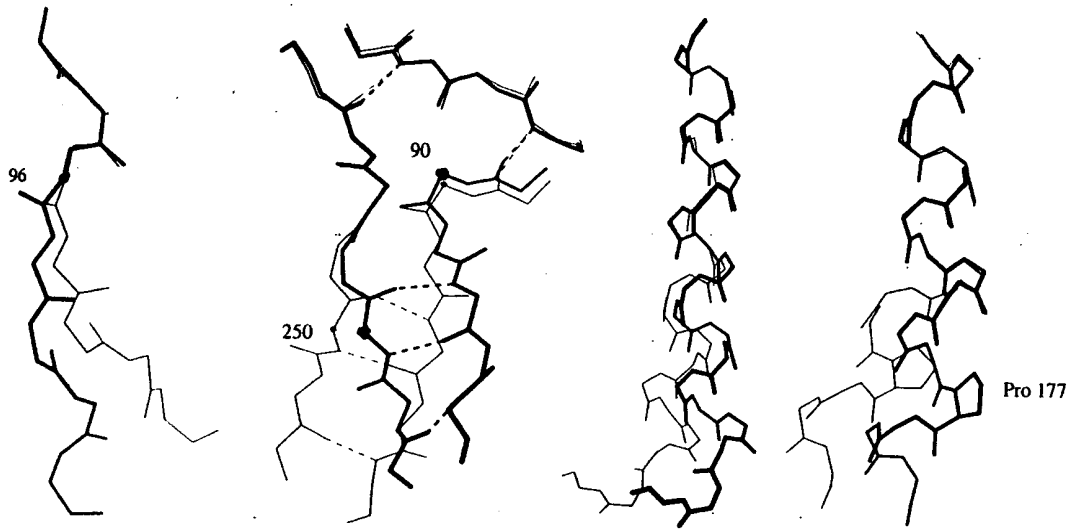

FIGURE 2: Hinge motions in strands, sheets, and helices. (A, far left) A hinge in lactate dehydrogenase is an example of a isolated hinge in a strand. Changes in two torsion angles ($\Delta\phi(96) = 36°$ and $\Delta\phi(97) = 40°$) are responsible for rotating the polypeptide chain ~35°. (B, middle left) The hinges in lactoferrin are an example of the coupling of two simple hinges together in a sheet. The hinges move through three large torsion angle changes, and the rotation axes for these torsion angle changes are inclined less than 20° with respect to the axis of the overall motion. (In the strand on the left $\Delta\psi(250) = -33°$ and $\Delta\phi(249) = 30°$; in the strand on the right $\Delta\psi(90) = 49°$.) Small conformational changes in adjacent residues help maintain the integrity of the β-sheet structure. As evident in Figure 6, the hinges have few main-chain packing constraints on them. (C, middle right) The interdomain helix in lysozyme is an example of a bending helix. It bends through the coordinated action of eight torsion angle changes between 9° and 15°, shifting the Cα atom at the C-terminal end of the helix by 3.3 Å. (D, far right) The helix linking the two domains in ADK is an example of a kinking helix. A torsion angle change in the residue three before Pro 177 ($\Delta\phi = -53°$) causes the helix to deform in a direction perpendicular to its original kink.

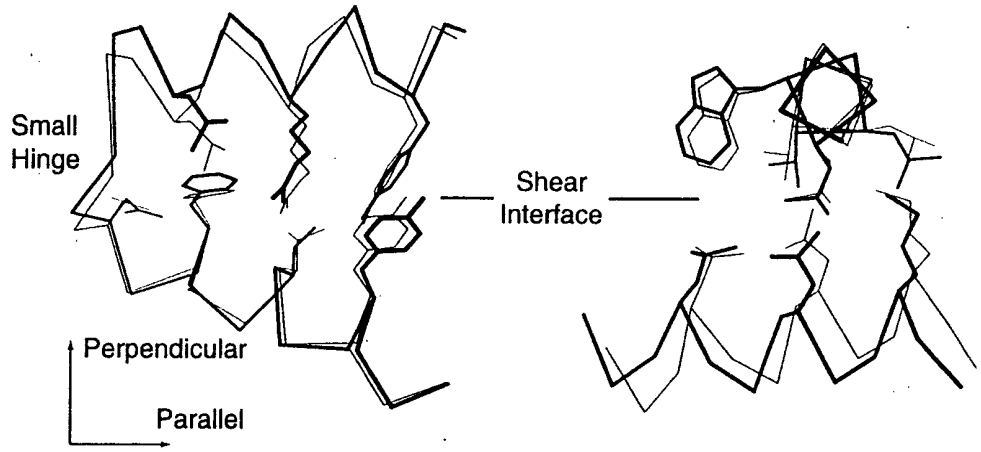

FIGURE 3: Shear motions involve interfaces. Two examples taken from citrate synthase show helix–helix interfaces undergoing a shear motion. The two labeled axes show the direction of parallel and perpendicular motion at an interface. (Left) The QP helix–helix interface illustrates how small hinges in linking peptides function in shear motions. Helix Q shifts 1.4 Å and rotates 13° relative to helix P. (Right) The NQ helix–helix interface shows a crossed-helix packing and a slightly larger motion than at the QP interface. Helix N shifts 1.8 Å and rotates 11° relative to helix Q. There are many close-packed side chains forming the N–Q interface, which just rock slightly in the shear motion.

changes in side-chain torsion angles. They keep the same overall rotamer configuration and move among conformational states of nearly the same energy without crossing large energy barriers. Occasionally, they may change to a different rotamer conformation (*i.e.*, to a different local minimum) with large rotations (>100°). (2)The main chain of each segment in a shear motion does not deform appreciably. In the case of helices, the root mean square difference in the positions of their main-chain atoms in the open and closed forms is typically 0.15–0.25 Å; for loops the difference is slightly larger. This rigidity, combined with "rocking" movements of side chains, implies that the interface itself shears. (3) The segments shift and rotate relative to each other by no more than 2 Å and 15°, amounts likely to be the limits of low-energy conformational adjustments. Except at very small interfaces, larger movements than these require the combination of several shear motions.

These characteristics were initially deduced from the analysis of protein crystal structures (Chothia *et al.*, 1983; Lesk & Chothia, 1984). A similar, and in some ways more detailed, picture of shear motions has recently emerged through physical studies and computational simulations (Elber & Karplus, 1987; Rojewska & Elber, 1990; Frauenfelder *et al.*, 1991).

SHEAR AND HINGE MOTIONS UNDERLIE DOMAIN-MOTION MECHANISMS

The characteristics of the two basic mechanisms of protein flexibility, hinge and shear motions, are summarized in Figure 1 and Table 1. These two mechanisms constitute a repertoire of conformational changes that can be used in a great variety of protein motions. Here we describe their use in the motions of whole protein domains, *i.e.*, in the relative motion of discrete linked units that consist, in most cases, of at least 100 residues.

Hinge and shear mechanisms are also involved in the motion of small protein fragments, for example, when individual loops or helices move relative to each other. In Table 2 we summarize the current crystallographic evidence for hinge and shear mechanisms in both domain motions and smaller motions. It is important to realize that hinge and shear motions are ideal paradigms for describing large domain motions. A real domain motion will often have a combination of both motions, *i.e.*, hinges in one part of the protein and shearing interfaces elsewhere. Nevertheless, many domain motions can be described as occurring predominantly by a hinge or a shear mechanism.

As shown in Figure 1, proteins that have a predominantly hinged domain motion usually have two domains connected by linking hinge regions that are relatively unconstrained by packing. A few large torsion angle changes are sufficient to produce almost the whole domain motion. The rest of the protein rotates essentially as a rigid body, with the axis of the overall rotation passing through the linking hinge regions.

Since an individual shear motion is small, a single one is usually not sufficient to produce a large domain motion. Usually, a number of shear motions combine to give a large effect—in a similar fashion to each block in a stack sliding slightly to make the whole stack lean considerably. (The peptides that link the shearing segments have small main-chain torsion angle changes to accommodate the relative movements.)

Proteins with shear motions tend to have certain architectural features. First, they often have layered architectures with one layer sliding over another. Second, though shear motions have been found at many different interfaces (*i.e.*, helix–helix, sheet–helix, loop–sheet, and loop–helix), helix–helix interfaces are most commonly used. The helices involved in shear motions are usually crossed. That is, they are usually oriented in a more perpendicular than parallel fashion (interhelical angle 60°–90°). Such crossed geometries are unusual in that helix–helix packings tend to be more parallel. Crossed helices will obviously have a smaller and more accommodating interface than parallel helices, and this is perhaps the reason for their preferential involvement in shear motions.

Table 2A lists all instances of crystallographically resolved domain motion, *i.e.*, proteins that have been solved in two or more conformations. With the notable exception of the immunoglobulins, almost all large domain motions can be understood in terms of hinge and shear motions. Table 2B lists proteins for which a domain closure mechanism can be inferred. The structures of these proteins have been determined in only one conformation. However, each has a structure similar to that of a protein with a well-characterized domain motion, *i.e.*, one listed in Table 2A, and is expected to move using the same mechanisms.

EXAMPLES OF SHEAR DOMAIN MOVEMENTS

(A) Citrate Synthase. Citrate synthase is one of the clearest examples of a domain closure occurring through shear motions. The molecule is a dimer, and each monomer comprises a large domain, containing 15 helices, and a small domain, containing five helices, with the active site cleft between them (Figure 4). The domain closure involves the small domain closing over the large one, burying the substrates in the active site (Remington *et al.*, 1982). An extensive interface between the large and small domains prevents closure taking place through a hinge mechanism. As shown in Figure 4, closure is produced by the summation of many small shear motions

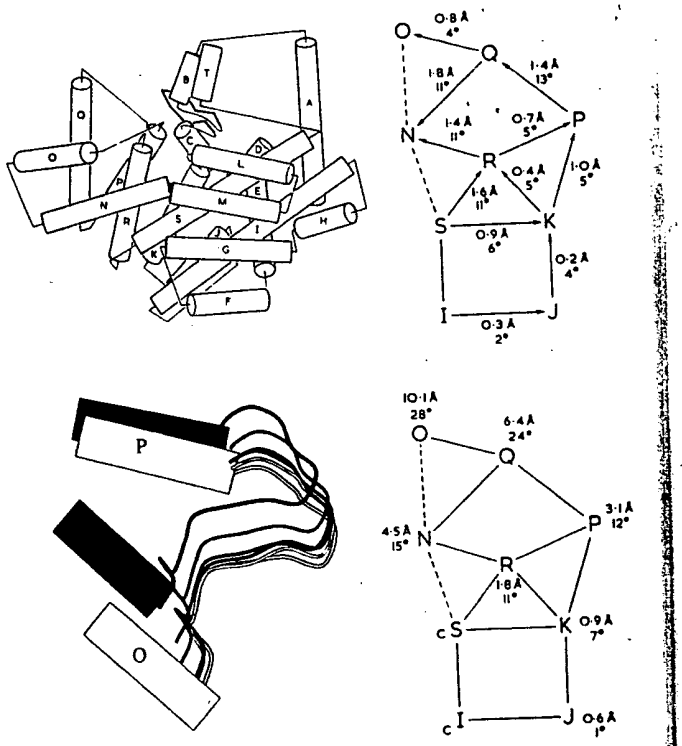

FIGURE 4: Shear motions in citrate synthase. (Top left) Cartoon of one subunit of citrate synthase. α-Helices are represented by cylinders. The small domain contains helices N, O, P, Q, and R. (top right) Schematic showing the relative movements of the principal helices in citrate synthase. [This figure is adapted in part from Lesk and Chothia (1984).] Each helix is represented by its letter, and the lines indicate the existence of helix–helix packings in both the open and closed forms. The shifts and rotations show local changes in the positions of pairs of packed helices (i.e., the movement in one helix in a pair relative to the other). (Bottom right) The overall effect of the helix movements. The same conventions as in the top right schematic apply, but the shifts and rotations shown now are those required to superimpose equivalent pairs of helices after the open and closed forms have been superimposed on the core of the large domain. Many small motions add up to shift helix O by 10.1 Å and rotate it by 28°. (Bottom left) Incremental motion in shear domain closure is shown by Cα traces of the OP loop: black is the apo form; white, the holo form; gray, the cumulative effect of motion over the K, P, and then Q helix–helix interfaces. (The apo form was fit to the holo form, first on the core and then on the K, P, and Q helices.)

between pairs of packed helices (Lesk & Chothia, 1984). The overall motion results in a helix on the far side of the small domain shifting by 10 Å and rotating by 28°, thereby moving an adjacent loop over the active site. Each local shear motion involves one helix moving relative to a neighboring helix by main-chain rotations and shifts of up to 13° and 1.8 Å. To a good approximation, the main chain of each helix moves without deformation as a rigid body. The shear motions are facilitated by small deformations in the loops linking the helices.

There are over 50 distinct helix–helix interfaces in the citrate synthase dimer. Depending on the angle between neighboring helices, these interfaces can be categorized as having roughly parallel helices, roughly perpendicular ones, or orientations in between. The interfaces between many of the moving helices tend to be roughly perpendicular, or "crossed", while the helices that are relatively motionless tend to have a more parallel packing.

(B) Aspartate Amino Transferase. In citrate synthase the domain closure is the cumulative result of many shear motions. In aspartate amino transferase (AAT) the domain motion is mainly the result of just two shear motions, which occur in

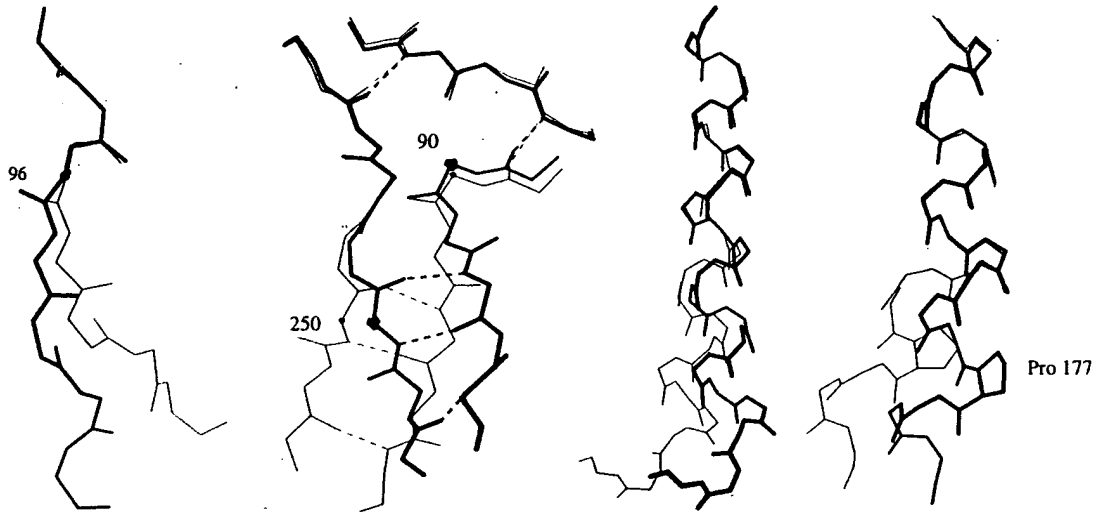

FIGURE 2: Hinge motions in strands, sheets, and helices. (A, far left) A hinge in lactate dehydrogenase is an example of a isolated hinge in a strand. Changes in two torsion angles ($\Delta\phi(96) = 36°$ and $\Delta\phi(97) = 40°$) are responsible for rotating the polypeptide chain ~35°. (B, middle left) The hinges in lactoferrin are an example of the coupling of two simple hinges together in a sheet. The hinges move through three large torsion angle changes, and the rotation axes for these torsion angle changes are inclined less than 20° with respect to the axis of the overall motion. (In the strand on the left $\Delta\psi(250) = -33°$ and $\Delta\phi(249) = 30°$; in the strand on the right $\Delta\psi(90) = 49°$.) Small conformational changes in adjacent residues help maintain the integrity of the β-sheet structure. As evident in Figure 6, the hinges have few main-chain packing constraints on them. (C, middle right) The interdomain helix in lysozyme is an example of a bending helix. It bends through the coordinated action of eight torsion angle changes between 9° and 15°, shifting the Cα atom at the C-terminal end of the helix by 3.3 Å. (D, far right) The helix linking the two domains in ADK is an example of a kinking helix. A torsion angle change in the residue three before Pro 177 ($\Delta\phi = -53°$) causes the helix to deform in a direction perpendicular to its original kink.

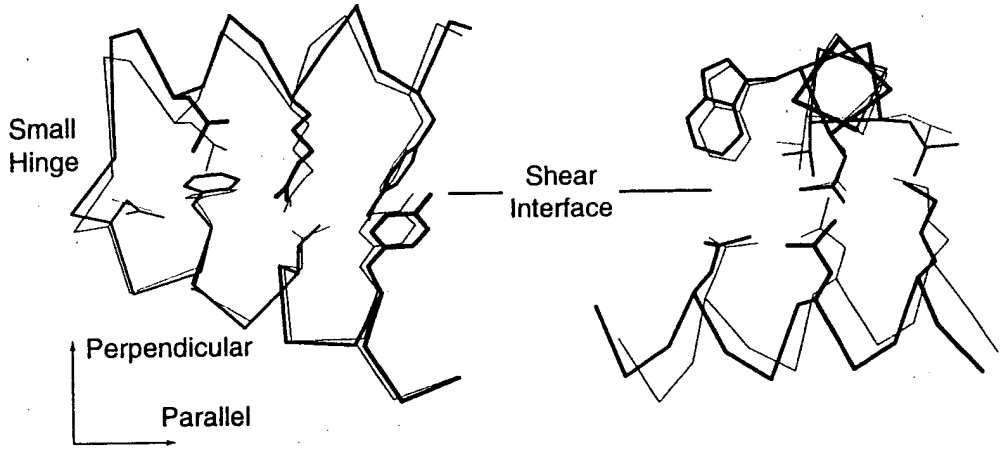

FIGURE 3: Shear motions involve interfaces. Two examples taken from citrate synthase show helix–helix interfaces undergoing a shear motion. The two labeled axes show the direction of parallel and perpendicular motion at an interface. (Left) The QP helix–helix interface illustrates how small hinges in linking peptides function in shear motions. Helix Q shifts 1.4 Å and rotates 13° relative to helix P. (Right) The NQ helix–helix interface shows a crossed-helix packing and a slightly larger motion than at the QP interface. Helix N shifts 1.8 Å and rotates 11° relative to helix Q. There are many close-packed side chains forming the N–Q interface, which just rock slightly in the shear motion.

changes in side-chain torsion angles. They keep the same overall rotamer configuration and move among conformational states of nearly the same energy without crossing large energy barriers. Occasionally, they may change to a different rotamer conformation (*i.e.*, to a different local minimum) with large rotations (>100°). (2)The main chain of each segment in a shear motion does not deform appreciably. In the case of helices, the root mean square difference in the positions of their main-chain atoms in the open and closed forms is typically 0.15–0.25 Å; for loops the difference is slightly larger. This rigidity, combined with "rocking" movements of side chains, implies that the interface itself shears. (3) The segments shift and rotate relative to each other by no more than 2 Å and 15°, amounts likely to be the limits of low-energy conformational adjustments. Except at very small interfaces, larger movements than these require the combination of several shear motions.

These characteristics were initially deduced from the analysis of protein crystal structures (Chothia *et al.*, 1983; Lesk & Chothia, 1984). A similar, and in some ways more detailed, picture of shear motions has recently emerged through physical studies and computational simulations (Elber & Karplus, 1987; Rojewska & Elber, 1990; Frauenfelder *et al.*, 1991).

SHEAR AND HINGE MOTIONS UNDERLIE DOMAIN-MOTION MECHANISMS

The characteristics of the two basic mechanisms of protein flexibility, hinge and shear motions, are summarized in Figure 1 and Table 1. These two mechanisms constitute a repertoire of conformational changes that can be used in a great variety of protein motions. Here we describe their use in the motions of whole protein domains, *i.e.*, in the relative motion of discrete linked units that consist, in most cases, of at least 100 residues.

Hinge and shear mechanisms are also involved in the motion of small protein fragments, for example, when individual loops or helices move relative to each other. In Table 2 we summarize the current crystallographic evidence for hinge and shear mechanisms in both domain motions and smaller motions. It is important to realize that hinge and shear motions are ideal paradigms for describing large domain motions. A real domain motion will often have a combination of both motions, *i.e.*, hinges in one part of the protein and shearing interfaces elsewhere. Nevertheless, many domain motions can be described as occurring predominantly by a hinge or a shear mechanism.

As shown in Figure 1, proteins that have a predominantly hinged domain motion usually have two domains connected by linking hinge regions that are relatively unconstrained by packing. A few large torsion angle changes are sufficient to produce almost the whole domain motion. The rest of the protein rotates essentially as a rigid body, with the axis of the overall rotation passing through the linking hinge regions.

Since an individual shear motion is small, a single one is usually not sufficient to produce a large domain motion. Usually, a number of shear motions combine to give a large effect—in a similar fashion to each block in a stack sliding slightly to make the whole stack lean considerably. (The peptides that link the shearing segments have small main-chain torsion angle changes to accommodate the relative movements.)

Proteins with shear motions tend to have certain architectural features. First, they often have layered architectures with one layer sliding over another. Second, though shear motions have been found at many different interfaces (*i.e.*, helix–helix, sheet–helix, loop–sheet, and loop–helix), helix–helix interfaces are most commonly used. The helices involved in shear motions are usually crossed. That is, they are usually oriented in a more perpendicular than parallel fashion (interhelical angle 60°–90°). Such crossed geometries are unusual in that helix–helix packings tend to be more parallel. Crossed helices will obviously have a smaller and more accommodating interface than parallel helices, and this is perhaps the reason for their preferential involvement in shear motions.

Table 2A lists all instances of crystallographically resolved domain motion, *i.e.*, proteins that have been solved in two or more conformations. With the notable exception of the immunoglobulins, almost all large domain motions can be understood in terms of hinge and shear motions. Table 2B lists proteins for which a domain closure mechanism can be inferred. The structures of these proteins have been determined in only one conformation. However, each has a structure similar to that of a protein with a well-characterized domain motion, *i.e.*, one listed in Table 2A, and is expected to move using the same mechanisms.

EXAMPLES OF SHEAR DOMAIN MOVEMENTS

(A) Citrate Synthase. Citrate synthase is one of the clearest examples of a domain closure occurring through shear motions. The molecule is a dimer, and each monomer comprises a large domain, containing 15 helices, and a small domain, containing five helices, with the active site cleft between them (Figure 4). The domain closure involves the small domain closing over the large one, burying the substrates in the active site (Remington *et al.*, 1982). An extensive interface between the large and small domains prevents closure taking place through a hinge mechanism. As shown in Figure 4, closure is produced by the summation of many small shear motions

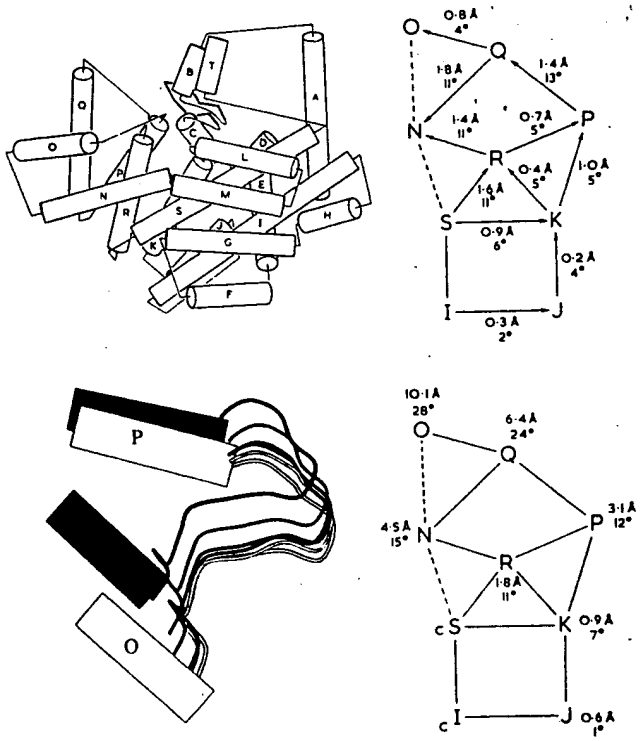

FIGURE 4: Shear motions in citrate synthase. (Top left) Cartoon of one subunit of citrate synthase. α-Helices are represented by cylinders. The small domain contains helices N, O, P, Q, and R. (top right) Schematic showing the relative movements of the principal helices in citrate synthase. [This figure is adapted in part from Lesk and Chothia (1984).] Each helix is represented by its letter, and the lines indicate the existence of helix–helix packings in both the open and closed forms. The shifts and rotations show local changes in the positions of pairs of packed helices (i.e., the movement in one helix in a pair relative to the other). (Bottom right) The overall effect of the helix movements. The same conventions as in the top right schematic apply, but the shifts and rotations shown now are those required to superimpose equivalent pairs of helices after the open and closed forms have been superimposed on the core of the large domain. Many small motions add up to shift helix O by 10.1 Å and rotate it by 28°. (Bottom left) Incremental motion in shear domain closure is shown by Cα traces of the OP loop: black is the apo form; white, the holo form; gray, the cumulative effect of motion over the K, P, and then Q helix–helix interfaces. (The apo form was fit to the holo form, first on the core and then on the K, P, and Q helices.)

between pairs of packed helices (Lesk & Chothia, 1984). The overall motion results in a helix on the far side of the small domain shifting by 10 Å and rotating by 28°, thereby moving an adjacent loop over the active site. Each local shear motion involves one helix moving relative to a neighboring helix by main-chain rotations and shifts of up to 13° and 1.8 Å. To a good approximation, the main chain of each helix moves without deformation as a rigid body. The shear motions are facilitated by small deformations in the loops linking the helices.

There are over 50 distinct helix–helix interfaces in the citrate synthase dimer. Depending on the angle between neighboring helices, these interfaces can be categorized as having roughly parallel helices, roughly perpendicular ones, or orientations in between. The interfaces between many of the moving helices tend to be roughly perpendicular, or "crossed", while the helices that are relatively motionless tend to have a more parallel packing.

(B) Aspartate Amino Transferase. In citrate synthase the domain closure is the cumulative result of many shear motions. In aspartate amino transferase (AAT) the domain motion is mainly the result of just two shear motions, which occur in

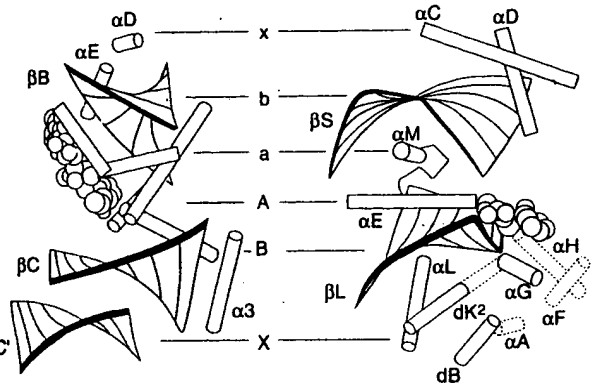

FIGURE 5: XBAabx layering in hexokinase and other proteins. XBAabx layering [see Examples of Shear Domain Movements (C)] is shown graphically by schematics of GAPDH (left) and hexokinase (right). Helices are drawn as narrow cylinders (radius 1.0 Å); sheets are represented as sheets as opposed to collections of strands; and substrates are drawn in CPK representation. (GAPDH is shown in its closed form with its actual ligand. Hexokinase is shown with the inhibitor *o*-toluoylglucosamine.)

perpendicular directions (McPhalen *et al.*, 1992). AAT has an active site situated between a large and a small domain, and on substrate binding the small domain closes over the active site. The major shear motion involves a 13° rotation of the core of the small domain relative to the large one. A secondary shear motion moves a helix on one side of the small domain in a direction perpendicular to both the interdomain interface and the direction of the other shear motion. With a 1.2-Å shift and a 10° rotation, it "drops down" to cover the active site.

The shear motions in AAT are facilitated by a hinge motion in a long interdomain helix. This helix is kinked by 17° in the open form and changes its kink angle by 12° on closure.

(C) Glyceraldehyde-3-phosphate Dehydrogenase, Alcohol Dehydrogenase, and Hexokinase. In the previous two examples, domain closure involved motions spread throughout a domain. Here we describe three examples where the major shear motion occurs at the interface between the domains with subsidiary motions on one or both sides of this region.

Because the enzymes hexokinase, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), and alcohol dehydrogenase (ADH) share many common architectural features, their domain movements proceed through similar mechanisms (see Table 2 for detailed references). These enzymes have three moving layers in one domain that shift relative to three rigid layers in the other domain. This distinctive layering pattern is of the form XBAabx, where a, b, and x are the three moving layers and X, B, and A are the rigid layers (Figure 5). The interface between the two middle layers is where the major shear motion occurs. One layer of helices from the mobile domain (a) slides over a layer of helices from the motionless domain (A). Helices in these two layers, which in a sense form gears upon which the domains slide, are often crossed, as is dramatically illustrated for the case of hexokinase in Figure 5. Near the a and A layer helices, the ligand binds in the interdomain cleft. Packed onto either side of the central layers of helices (a and A) are sheets (b and B) from the mobile and motionless domains, respectively. The mobile sheet (b) forms a second moving layer, which slides over the helices, and packed onto the other face of this sheet is a third layer (x) which moves with the sheet. Symmetrical to this third moving layer (x), a third motionless layer (X) is packed onto one side of the static sheet (B). (Layer x is made up of helices in hexokinase and GAPDH and of helices and a sheet in ADH, and layer X is made up of helices in hexokinase and ADH and of a sheet from another subunit in GAPDH.)

In addition to its shear motion, ADH also has two well-defined hinge points (Eklund *et al.*, 1981; Colonna-Cesari *et al.*, 1986).

As discussed in Table 2B, a number of other proteins have XBAabx architectures similar to those of hexokinase, GAPDH, and ADH but have not yet been solved in multiple conformations. These proteins include phosphoglycerate kinase (PGK), actin, and heat-shock protein. There is experimental evidence that these proteins may undergo domain movements [*e.g.*, for PGK, see Mas *et al.* (1987, 1988)], and they would be expected to use mechanisms similar to those of hexokinase, GAPDH, and ADH. Moreover, a model-building study done on PGK (Blake *et al.*, 1986) predicts that the domain movement will involve the shearing of the two central helices, a conclusion similar to that implied by our comparisons.

(D) trp Repressor. In the previous sections we described examples of domains closing around substrates. In the *trp* repressor, the binding of a ligand stabilizes a more open conformation. The *trp* repressor is a small protein that regulates three operons involved in the synthesis of tryptophan. It is a dimer, and each subunit contains six helices, divided between two domains. The central core of the molecule is formed from four helices from each subunit. On either side of this core, helix–turn–helix motifs form two symmetrically arranged DNA "reading head" domains. Between the central core and the reading-head domains, there are two binding sites for L-tryptophan, which need to be filled for *trp* repressor to recognize DNA (Zhang *et al.*, 1987). Comparison of the holo and apo forms of the repressor (Lawson *et al.*, 1988) shows that the binding of L-tryptophan shifts Cα atoms in the reading head domain by up to 4 Å. These shifts are produced by separate shear motions of the two helices in the reading-head domain (0.75–1.5 Å, 5–20°). These helix motions move the reading-head domains further apart than they are in the apo form so they are correctly separated to bind DNA.

EXAMPLES OF HINGED DOMAIN MOVEMENTS

(A) Tomato Bushy Stunt Virus. An example of a very simple hinge motion is found in the coat protein of tomato bushy stunt virus (Olsen *et al.*, 1983). This spherical virus contains 180 subunits arranged with icosahedral symmetry on a $T = 3$ lattice. Each subunit, in turn, contains two major domains, the shell (S) and projection (P) domains, that are linked by a peptide in an extended conformation. The symmetry of the virus requires each subunit to fit into one of three different packing environments. One of the principal mechanisms for accommodating the different environments is a relative movement of the two domains by ~22°. This movement involves a simple hinge in the peptide connecting the S and P domains (Olsen *et al.*, 1983).

(B) Calmodulin. Like the TBSV coat protein, the motion in calmodulin involves a single deformation. The unligated form of calmodulin contains two globular domains, connected by a long helix (Babu *et al.*, 1985). NMR and X-ray structures of ligated calmodulin show the molecule binding to peptide helices with different sequences and the two domains closing around the peptide far enough to make contact with each other (Ikura *et al.*, 1992; Meador *et al.*, 1992, 1993). As discussed above [The Intrinsic Flexibility of Proteins (A) (3)], in this motion, the long interdomain helix, which is known to have only marginal stability in solution (Ikura *et al.*, 1992),

Table 2: Proteins That Undergo Domain Movements

(A) Proteins for which open and closed conformations are known [a b]

(i) Domain motion is predominantly shear

Citrate Synthase [c]	1CTS 3CTS	Remington *et al.*, 1982; Lesk & Chothia, 1984	Shear motions at many helix-helix interfaces shift mainchain atoms up to 10 Å
Aspartate Amino Transferase (AAT) [c]	9AAT 1AMA	McPhalen *et al.*, 1992	Shear motion at 2 interfaces combined with hinge in a kinked helix.
Trp Repressor [c]	1WRP 2WRP 3WRP	Lawson *et al.*,1988	Shear motion between 2 helices adjusts position of helix-turn-helix reading head domain to enable it to bind DNA
Hexokinase [c]	2YHX 1HKG	Bennett & Steitz, 1978, 1980; Lesk & Chothia, 1984	Shear motion with XBAaba layering. Prominent crossed helices at interdomain interface.
Glyceraldehyde-3-phosphate Dehydrogenase (GAPDH) [c]	1GD1 2GD1	Skarzynski & Wonacott, 1988	Shear motion with XBAaba layering.
Alcohol Dehydrogenase (ADH) [c]	8ADH 6ADH	Eklund *et al.*,1981; Colonna-Cesari *et al.*, 1986	Shear motion with XBAaba layering and 2 hinges.
Endothiapepsin	4APE 5ER2	Sali *et al.*, 1989; 1992	Small shearing motion at 1 interface between domains (17° rotation and 1 Å displacement)

(ii) Domain motion is predominantly hinge

Tomato Bushy Stunt Virus (TBSV) Coat Protein [c e]	2TBV	Olson *et al.*, 1983	1 interdomain linkage,1 hinge, ~22° rotation.
Lactoferrin [c]	1LFH 1LFG	Anderson *et al.*, 1990; Gerstein *et al.*, 1993b	2 interdomain linkages, 2 hinges (in a β-sheet), 53° rotation. See-saw between two interfaces.
Maltodextrin Binding Protein (MBP) [c]	1OMF 2MBP	Sharff *et al.*, 1992;	3 interdomain linkages, 3 hinges, 35° rotation.
Lysine/Arginine/Ornithine (LAO) binding protein [c]	1LST	Oh *et al.*, 1993	2 interdomain linkages, 2 hinges, 52° rotation.
T4 lysozyme mutants: Ile3→Pro & Met6→Ile [c]	1L96 1L97	Dixon *et al.*, 1992; Faber & Matthews, 1991	2 hinges, at either end of interdomain helix, produce rotations up to 32°.
Adenylate Kinase (ADK) [c g]	1AK3 1AKE	Schulz *et al.*, 1990; Gerstein *et al.*, 1993a	2 interdomain linkages and 4 hinges (one involves kinking helix). 60° rotation from 1st pair of hinges, 30° from 2nd pair, 90° total.
Catabolite Gene Activator Protein (CAP) [e]	3GAP	Weber & Steitz, 1987	1 interdomain linkage and 1 hinge. Comparison of subunits in the dimer reveals that the small domain has rotated ~30° closer to the large domain in one subunit.
cAMP-dependent Protein Kinase (catalytic domain) [c d]	1ATP 1APM	Karlsson *et al.*, 1993	1st set of hinges, involving 3 interdomain linkages, produces 12° rotation of domain cores (with ~3 Å shift). 2nd set of hinges produces further 6° rotation of a loop. 1 shearing interface between domains.
Calmodulin [c]	1CLL 4CLL 2BBM	Ikura *et al.*, 1992; Meador *et al.*, 1992, 1993	1 interdomain linkage, 1 hinge, ~150° rotation. Hinge involves long helix splitting into 2 helices (inclined at ~100°) with strand in between.
Glutamate Dehydrogenase		Stillman *et al.*, 1993	13° rotation of 1 domain relative to other

(iii) Domain motion is not predominantly a hinge or shear mechanism

Immunoglobulins [c h]	2FB4 1MCP	Bennett & Huber, 1984; Lesk & Chothia, 1988;	Hinge motion in linking peptides. Ball & socket joint forms interface between domains. Range of rotations up to 50° allowed.
Serpins	5API 1OVA	Loebermann *et al.*, 1984 Engh *et al.*, 1990 Stein & Chothia, 1991 Mottonen *et al.*, 1992	Translation at a helix-sheet interface results in the transformation of the tertiary structure by insertion of strand into sheet.

(iv) Domain motion can not be fully classified at present [f]

HI V-1 Reverse Transcriptase	1HMI 1HVT	Kohlstaedt *et al.*, 1992; Jacobo-Molina *et al.*, 1993	Comparison of subunits shows very large rearrangement of 2 of the 4 domains which is accomodated by changes in loops and by unfolding of small 3 stranded β-sheet.
TATA-box Binding Protein (TBP) [e]	1TBP	Kim *et al.*, 1993a, 1993b; Chasman *et al.*, 1993	Twisting of a central sheet moves 2 domains ~10°.
Themolysin, Elastase, neutral proteases	1EZM 4TMN	Holland *et al.*, 1992; Thayer *et al.*, 1991	Bending interdomain helix
Elongation Factor Tu (EF-Tu) [d]	1ETU	Berchtold *et al.*, 1993; Kjeldgaard *et al.*, 1993	Internal loop movements similar to those in *ras* protein (below) lead to large domain rearragements (90° rotation, 40Å shifts)

Table 2: (Continued)

(B) Proteins for which only one conformation is known

(I) Domain motion is predominantly shear

Phosphoglycerate Kinase (PGK) [c]	3PGK	Harlos *et al.*, 1992	Similar to hexokinase (XBAabx layering)
Heat Shock Protein	1HSC	Flaherty *et al.*, 1990	Similar to hexokinase (XBAabx layering)
Actin	1ATN	Kabsch *et al.*, 1990; Flaherty *et al.*, 1991 [i]	Similar to hexokinase (XBAabx layering)
Aspartic Proteases, besides endothiapepsin: Peniclopepsin, Rhizopuspepsin, Chymosin, Porcine Pepsin	2APP 2APR 2PEP 3CMS 1PSG	Sali *et al.*, 1992	Similar to endothiapepsin

(II) Domain motion is predominantly hinge

Sulfate & Phosphate Binding Proteins	1SBP 1ABH	Luecke & Quiocho, 1990; Pflugrath & Quiocho, 1988	Similar to MBP & lactoferrin. These are group-II periplasmic binding proteins.
Arabinose, Leucine, & Galactose Binding Proteins	2LBP 2GBP 1ABP	Gilliland & Quiocho, 1981; Vyas *et al.*, 1988, 1991; Sack *et al.*, 1989a,b	Similar to MBP & lactoferrin. However, these are group-I periplasmic binding proteins and are not as similar as group-II ones (above) are.
Transferrins (N-terminal lobe)	1TFD	Sarra *et al.*, 1990	Similar to lactoferrin
Guanylate Kinase (GDK)	1GKY	Stehle & Schulz, 1990	Similar to ADK
Porphobilinogen Deaminase	1PDA	Louie *et al.*, 1992	Domains 1 and 2 similar to lactoferrin

(III) Domain motion can not be classified at present [f]

Myosin		Rayment *et al.*, 1993	Closure of a nucleotide-binding cleft, with similarities to that of ADK, hypothesized to produce movements > 50 Å
Transducin-α		Noel *et al.*, 1993	Similar movements to EF-Tu and ras expected

(C) Proteins known in two conformations which involve movements of fragments smaller than domains [a]

(I) Motion is predominantly shear

Insulin [d]	4INS	Chothia *et al.*, 1983	Helices shear by ~1.5 Å.
Thymidylate Synthase	3TMS 2TSC	Perry *et al.*, 1990; Montfort *et al.*, 1990	Small shear motion of helices packed onto central sheet.
Dihydrofolate Reductase (DHFR)	4DFR 5DFR	Bystroff *et al.*, 1991	Small (~3 Å) movement, shearing interface with hinges.

(II) Motion is predominantly hinge

Annexin V	1AVR 1RAN	Sopkova *et al.*, 1993; Concha, *et al.*, 1993	Large movements of 2 loops and end of a helix moves a buried trp residue 18Å to surface.
Lactate Dehydrogenase (LDH)	6LDH 1LDM	White *et al.*, 1976; Gerstein & Chothia, 1991	Loop closure with 2 hinges, one in helix, moves Cα atoms ~11 Å
Triose Phosphate Isomerase (TIM)	2YPI 3TIM 6TIM	Lolis & Petsko, 1990; Joseph *et al.*, 1990; Wirenga *et al.*, 1991	Loop closure with 2 hinges moves Cα atoms ~ 7Å
Enolase	3ENL 7ENL	Lebioda & Stec, 1991	Loop movements of ~7 Å
HIV-1 protease	4HVP 3HVP 5HVP	Miller *et al.*, 1989; Fitzgerald *et al.*, 1990	Two large loop regions, that together comprise one quarter of the structure, move Cα atoms ~ 7Å
Foot and mouth disease virus [d]	1BBT	Parry *et al.*, 1990	Comparing variants of virus shows movement of a surface loop
Triglyceride Lipase	1TGL 4TGL	Derewenda *et al.*, 1992;	2 hinges on either side of a helix move Cα atoms up to12 Å. In one hinge a residue changes from an extended to a helical conformation.
Isocitrate Dehydrogenase [d]	3ICD	Stoddard & Koshland, 1993	Loop movements of ~2 Å
Malate Dehydrogenase (MDH) [e]	4MDH	Birktoft *et al.*, 1989	Comparison of subunits shows a loop closure similar to LDH, moving atom Cα atoms up to 8 Å.
ras Protein	4Q21 6Q21	Milburn *et al.*, 1990; Sclichting *et al.*, 1990	2 loop movements move Ca atoms up to 10 Å (one movement includes helix attached to loop).

[a] When both open and closed forms are known, we refer to the papers that describe the structure comparisons. Further references to the individual open and closed structures can be found in these papers. [b] Allosteric proteins are not included because these proteins have motions that involve extensive repacking of interfaces (see Perutz, 1989 for a review). Such repacking involves high-energy conformational transitions distinctly different from the hinge and shear mechanisms. [c] Indicates proteins discussed in detail in the text. [d] Structures of 2 conformations have been solved but only 1 has been deposited in the Protein Data Bank. [e] Motion is evident in comparing different subunits in the asymmetric unit. Single data bank identifier applies for both forms. [f] It is not possible to classify some domain motions at present because full sets of coordinates or detailed analyses are not yet available. [g] ADK also has a shear motion when the first substrate, AMP, binds: i.e. in moving from the conformation of 3ADK to 1AK3, 3 helices with a crossed geometry shift 1-2 Å to rearrange the geometry of the nucleotide binding site slightly (Diederichs & Schulz, 1991). [h] Data bank indentifiers for only two of the many representative immunoglobulin stuctures are indicated. [i] This paper describes the structural similarity of actin and the heat-shock protein.

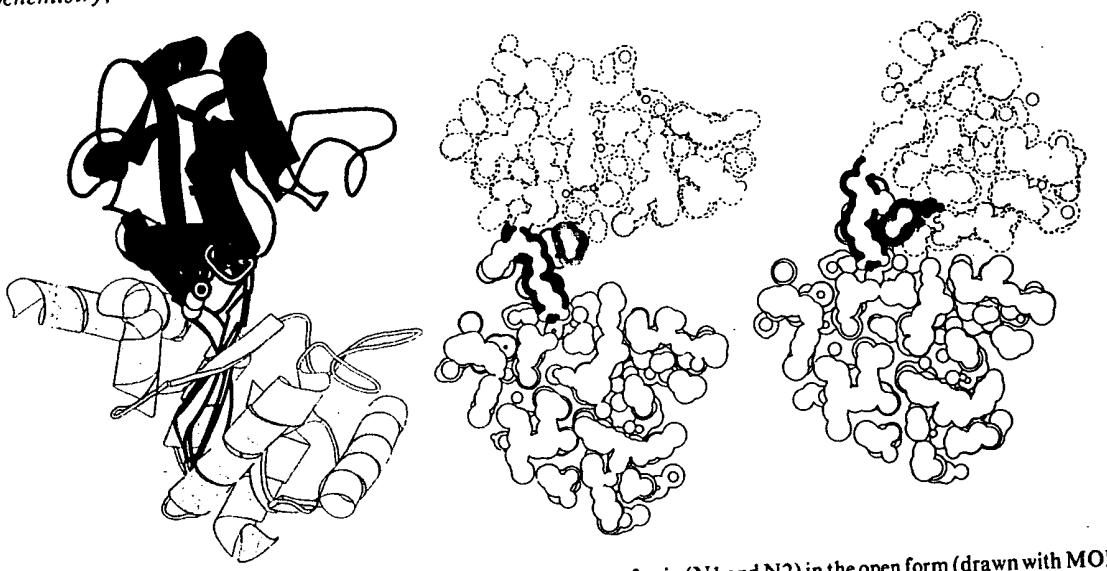

FIGURE 6: Hinge motion in lactoferrin. (Left) Cartoon of the two domains of lactoferrin (N1 and N2) in the open form (drawn with MOLSCRIPT; Kaulis, 1991). The origin of the rotation axis for the domain movements [see Examples of Hinged Domain Movements (A)] lies at the center of the figure. The view is down the rotation axis, which is indicated by a circle with a dot in it. N2 is shown in darker shading than N1, and the two antiparallel β-strands with the hinges are highlighted with bold lines. The Cα atoms of the residues with the largest movements (90 and 251) are indicated by empty circles. They lie in the middle of these strands and are very near the rotation axis in the open form. (Middle and right) Slices through the van der Waals envelope of the open and closed forms, respectively. N1 is shown by thin black lines; N2, by dashed black lines; the main-chain atoms of the hinge (89–92 and 249–252), by black shading; the side chain of Tyr 92, by a stippled gray line. Note the absence of packing constraints on the main-chain atoms of the hinge in the open form and the tight packing at the base of the hinge in the closed form.

partly unfolds to break into two helical segments connected by a hinge region in an extended conformation. The angle between the axes of the two helical segments is ~100°. As there is an additional twist around the helix axes, the total rotation of one domain relative to the other is more than 150°. Calmodulin can bind peptides with different sequences because of flexibility in the side chains that make contact with the peptide and by slightly shifting the relative placement of the domains through changes in the extent of the hinge region, which has consequently been dubbed "a variable expansion joint" (Meador *et al.*, 1993).

(C) T4 Lysozyme Mutants. Like calmodulin, two mutants of T4 lysozyme (Ile 3 → Pro and Met 6 → Ile) have a hinge motion involving a long interdomain helix. Crystals of these mutants grow in a number of different forms. Depending on the crystal form, their structures either are very similar to that of the wild type or differ from it by a *range* of rigid-body domain rotations up to 32° (Dixon *et al.*, 1992; Faber & Matthews, 1990). There are two main hinge points for the domain motion. They occur at the ends of the long helix that spans the domains. As discussed above, the second hinge involves small torsion angle changes spread throughout the C-terminal part of the helix (Figure 2C). As the location of the mutation is next to the hinge, the domain motion appears to be a consequence of the loss of close packing created by the mutation and is an example of hinged motion created by reducing the number of steric constraints.

(D) Lactoferrin and the Periplasmic Binding Proteins. Unlike the TBSV coat protein, lysozyme, and calmodulin, lactoferrin and the periplasmic binding proteins have two or three interdomain linkages, containing hinges. These proteins are examples of transport proteins that use domain closure to recognize and sequester small molecules.

Lactoferrin has two similar lobes, and each lobe, in turn, has two domains with an iron-binding site between them. Analyses of the open and closed forms of one of lobes give a detailed picture of the domain movements (Anderson *et al.*, 1990; Gerstein *et al.*, 1993b). Upon binding iron, the two domains move together, rotating 53° essentially as rigid bodies. The axis of rotation passes through the two β-strands linking the domains (Figure 6). As discussed above (Figure 2b), these strands contain distinct hinges, and as the rotation axes of the principal torsion angle changes are nearly parallel to the axis of the overall 53° rotation, the local motion in the hinges can be directly related to the overall domain closure.

The two domains make different packing contacts in the open and closed forms. In the open form the contacts are on one side of the hinges, and in the closed form they are on the other side. Pivoting about the hinges produces a seesaw motion between the two interfaces: when the domains close, residues in the interface on one side of the hinges become buried and close-packed, and residues on the other side become exposed. The situation is reversed on opening.

Lactoferrin shares a similar structure, topology, and binding site construction with the group II periplasmic binding proteins (Baker *et al.*, 1987). For two of these binding proteins, maltodextrin-binding protein and LAO-binding protein (Sharff *et al.*, 1992; Spurlino *et al.*, 1991; Oh *et al.*, 1993), structures have been determined for both the open and closed forms, and the mechanism of domain movement appears to be similar to that in lactoferrin. The domain motion in the maltodextrin-binding protein is a 35° rotation about an axis through the hinge region, and there are large, localized torsion angle changes in the three peptides linking the domains. The positions of two of the hinges are structurally equivalent to those of the lactoferrin hinges. In the LAO-binding protein there is a 52° rotation of the two domains, which involves only a few large torsion angle changes in a region structurally equivalent to the lactoferrin hinge.

(E) Adenylate Kinase. A more complex and extensive hinge motion is seen in the large variants of adenylate kinase. This enzyme has two nucleotide binding sites, and crystal structures have been solved with both sites, a single site, and no sites filled (Schultz *et al.*, 1974, 1990; Diederichs & Schulz, 1991; Müller & Schulz, 1992). The major conformational change,

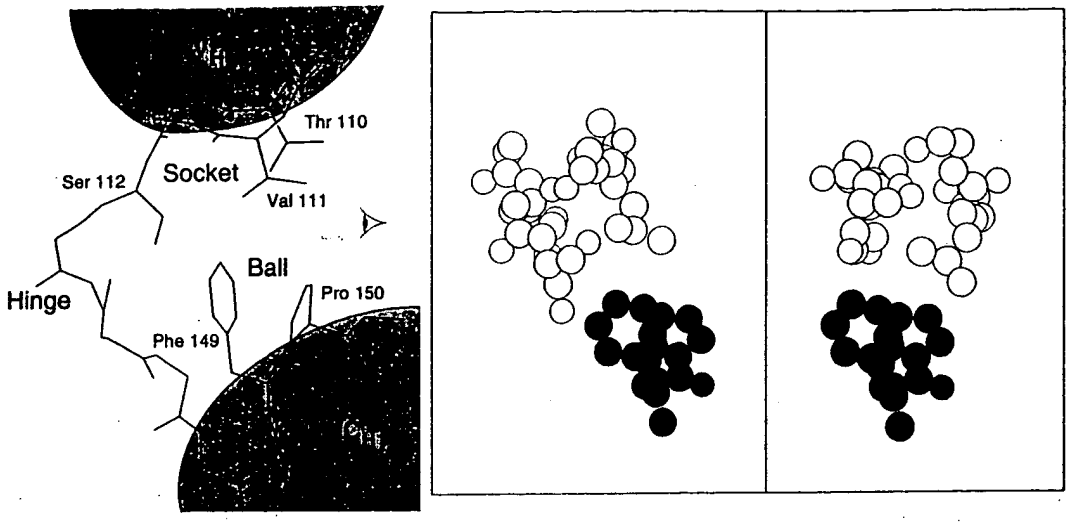

FIGURE 7: Ball-and-socket motion in the immunoglobulins. (Left) The conserved V_H–C_{H1} contacts and the switch (hinge) peptides. Three V_H residues (11, 110, and 112) form a "socket", and two C_{H1} residues, 149 and 150, form a "ball". The view is such that the motion of the V dimer relative to the C dimer is perpendicular to the page. (Middle and Right) The movement of the ball-and-socket joint. The five side chains in the joint are represented by spheres drawn at one-half van der Waals radius. White spheres indicate the socket, and black ones the ball. The orientation is roughly perpendicular to that in the left figure (see the eye symbol there). The middle figure shows the packing that occurs when the domains are fully extended (*i.e.*, 180° elbow angle), and the right figure shows the packing that occurs when the domains are close enough to be in contact (i.e., 135° elbow angle).

which occurs when the second substrate binds, rotates the smaller of the two domains ~90° relative to the larger one and shifts main-chain atoms up to 32 Å. The small and large domains are linked by two helices, and on closure, conformational changes take place in four hinges at the N and C termini of these linking helices (Gerstein *et al.*, 1993b). Two of these hinges have simple motions; a third hinge requires motion throughout an extended loop; and a fourth hinge (Figure 2D) occurs in the middle of a proline-kinked helix. The four hinges have few packing constraints on their main chain. One pair of hinges is responsible for one-third of the total rotation, and the other pair, for two-thirds.

(F) cAMP-Dependent Protein Kinase. Like ADK, the catalytic subunit of cAMP-dependent protein kinase has an elaborate multipart hinged motion, which involves at least five distinct hinges, split into the two sets. Containing two domains, one large and one small, the structure of the catalytic subunit has been solved in binary and ternary complexes with an inhibitory peptide and in an apo form (Knighton *et al.*, 1991; Karlsson *et al.*, 1993). In a comparison of the apo form with either complex form, the core of the small domain rotates ~12° relative to that of the large one. The small domain is principally connected to the large domain through three roughly parallel peptide linkages, which deform as hinges upon closure. In addition, through the deformation of two more hinges a loop in the small domain near the binding pocket rotates a further 6° down into the interdomain cleft. Partly because of the size of the interdomain cleft, which has to accommodate a 15-residue peptide, the protein kinase motion does not involve an extensive interdomain interface. There is, however, one helix in the small domain which moves in a shear fashion to maintain its contacts with the large domain throughout the motion.

THE BALL-AND-SOCKET MOTION IN THE IMMUNOGLOBULINS

The domain motion observed in the immunoglobulins involves, so far as is known at present, a unique combination of hinge and shear motions. In the immunoglobulins the V_L domain is linked by an extended peptide to the C_L domain, and V_H is similarly linked to C_{H1}. V_L and V_H pack together, as do C_L and C_{H1}. The V_L–V_H dimer can freely rotate, relative to the C_L–C_{H1} dimer, over a range of ~50° in a manner described as "elbow motion".

Elbow motion involves localized deformations in the two peptides that link the V and C dimers (Bennett & Huber, 1984). These deformations are similar to those in the hinged domain closures described in the previous sections. However, the elbow motion also involves an unusual type of shear motion: two large residues in C_{H1}, a Pro and a Phe, pack closely together, forming a "ball", and three residues in V_H spread out as part of a β-sheet, forming a "socket" (Figure 7). The three V_H and two C_{H1} residues are packed together and move relative to each other in a manner similar to a socket moving over a ball (Lesk & Chothia, 1988).

Unlike the shear motions discussed above, which are characterized by close-packed interfaces of interdigitating sidechains, the ball-and-socket joint has a "smooth" interface, in which the side chains do not interdigitate. This interface facilitates motion over a wide range of relative orientations. It also permits greater flexibility than is found in shear motions: the socket residues can move up to 4.5 Å relative to those in the ball, rather than the 1.5–2.0-Å displacement usually found at an interface undergoing shear motion.

THE STABILITY OF THE CLOSED AND OPEN STATES

The evidence currently available suggests that the open and closed states are only slightly different in energy and at room temperature are in dynamic equilibrium. This small energy difference between the open and closed states is most directly suggested by the discovery that relatively weak crystal packing forces can stabilize the unliganded closed forms of lactoferrin and the binding proteins (Baker *et al.*, 1991; Sharff *et al.*, 1992, and references therein). It is also suggested by simulations of loop closure (Wade *et al.*, 1993, 1994).

The relative stabilities of the open and closed states depend on the presence or absence of the substrate. A likely

progression is that the substrate first binds to one domain, then thermal fluctuations bring the second domain into contact with it, and the newly formed contacts stabilize the closed conformation. The ability of a ligand to bind to a single domain has, in fact, been observed in transferrin (Lindley *et al.*, 1993). Inspection of the structures of liganded closed states invariably shows that the ligand makes numerous interlocking salt bridges, hydrogen bonds, and packing interactions with both domains (references in Table 2), and these interactions account for the stability and specificity of the closed state. Catalytic transformation of the substrate destroys, at least in part, the interactions made with the protein and so makes the open form more stable. The rate of domain movements, consequently, is governed to a degree by the catalytic efficiency of the protein. This may be particularly relevant to domain movements involved in locomotion.

The main function of the open form is to allow access to the active site. By itself, this function does not require the open form to have a unique conformation, as opposed to a range of conformations. Experimental evidence for a unique open form is sketchy and mixed. On the one hand, there is clear evidence that the open form has a unique conformation in certain proteins. As discussed above [Examples of Hinged Domain Movements (C)], in lactoferrin the interdomain interface formed in the open form appears to uniquely fix its conformation. Likewise, within particular species, AAT has the same open conformation in different crystal forms, which have very different intermolecular contacts (McPhalen *et al.*, 1992). On the other hand, there is also evidence that the open form of other proteins can have a range of conformations. T4 lysozyme has been found to have a number of different "open" conformations in various crystal forms (Faber & Matthews, 1991; Dixon *et al.*, 1992). The leucine/isoleucine/valine-binding protein has been solved in a "more-opened" form (Sharff *et al.*, 1992, and references therein). A variety of different orientations have been found for the two domains of *Escherichia coli* $NADP^+$-dependent glutamate dehydrogenase; this hexameric protein has been solved in a crystal form where all six subunits are in the asymmetric unit (D. Rice, personal communication).

Note that the crystallographic evidence relating to the uniqueness of end states must be treated with care as there is a possibility that the intermolecular contacts in the crystal may fix domains in orientations not preferred in solution. Also, crystallography tends to make one think in terms of discrete, rigid conformational states, which may be an erroneous model for open and closed conformations.

CONCLUSIONS

We have shown how hinge and shear motions, which constitute the repertoire of low-energy conformation changes available to proteins, can be combined to describe most of the known instances of domain movements. We emphasize the importance of the architecture of the interdomain interface in determining the relative mix of hinge and shear motions. While our hinge and shear mechanisms do not describe domain motions precisely enough for accurate energy calculations, they provide a conceptual framework for understanding complicated structural transformations and can be used as a guide for more quantitative formulations. As more data become available, the descriptions of hinge and shear mechanisms should be refined and extended so that they can be applied to the complex large-scale motions that occur in structures such as myosin (Rayment *et al.*, 1993).

An expanded and routinely updated version of Table 2 (a listing of protein structures that undergo conformational change) will be available electronically in plain text and hypertext forms. Use (i) anonymous ftp or WWW with URL "file://cb-iris.stanford.edu/pub/ProteinMovements/ProteinMovements.html", (ii) anonymous ftp to "al.mrc-lmb.cam.ac.uk" for filename "pub/ProteinMovements/ProteinMovements.html", or (iii) electronic mail to mbg@cb-iris.stanford.edu.

REFERENCES

Anderson, B. F., Baker, H. M., Norris, G. E., Rumball, S. V., & Baker, E. N. (1990) *Nature 344*, 784.

Anderson, C. M., Zucker, F. H., & Steitz, T. (1979) *Science 204*, 375.

Babu, Y. S., Sack, J. S., Greenhough, T. J., Bugg, C. E., Means, A. R., & Cook, W. J. (1985) *Nature 315*, 37.

Baker, E. N., Rumball, S. V., & Anderson, B. F. (1987) *Trends Biochem. Sci. 12*, 350.

Bennett, W. S., & Huber, R. (1984) *Crit. Rev. Biochem. 15*, 291.

Bennett, W. S., Jr., & Steitz, T. A. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 4848.

Bennett, W. S., Jr., & Steitz, T. A. (1980) *J. Mol. Biol. 140*, 211.

Berchtold, H., Reshetnikova, L., Reiser, C. O. A., Schirmer, N. K., Sprinzl, M., & Hilgenfeld, R. (1993) *Nature 365*, 126.

Birktoft, J. J., Rhodes, G., & Banaszak, L. J. (1989) *Biochemistry 28*, 6065.

Blake, C. C. F., Rice, D. W., & Cohen, F. E. (1986) *Int. J. Pept. Protein Res. 27*, 443.

Bystroff, C., & Kraut, J. (1991) *Biochemistry 30*, 2227.

Chasman, D. I., Flaherty, K. M., Sharp, P. A., & Kornberg, R. D. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 8174.

Chothia, C., Lesk, A. M., Dodson, G. G., & Hodgkin, D. C. (1983) *Nature 302*, 500.

Colonna-Cesari, F., Perahia, D., Karplus, M., Eklund, H., Brändén, C. I., & Tapia, O. (1986) *J. Biol. Chem. 261*, 15273.

Concha, N. O., Head, J. F., Kaetzel, M. A., Dedman, J. R., & Seaton, B. A. (1993) *Science 261*, 1321.

Derewenda, U., Brzozowski, A. M., Lawson, D. M., & Derewenda, Z. S. (1992) *Biochemistry 31*, 1532.

Diederichs, K., & Schulz, G. E. (1991) *J. Mol. Biol. 217*, 541.

Dixon, M. M., Nicholson, H., Shewchuk, L., Baase, W. A., & Matthews, B. W. (1992) *J. Mol. Biol. 227*, 917.

Eklund, H., Samaha, J. P., Wallen, L., Branden, C. I., Åkeson, Å., & Jones, T. A. (1981) *J. Mol. Biol. 146*, 561.

Elber, R., & Karplus, M. (1987) *Science 235*, 318.

Engh, R. A., Wright, H. T., & Huber, R. (1990) *Protein Eng. 3*, 469.

Faber, H. R., & Matthews, B. W. (1990) *Nature 348*, 263.

Fitzgerald, P. M. D., McKeever, B. M., VanMiddlesworth, J. F., Springer, J. P., Heimbach, J. C., Leu, C.-T., Herber, W. K., Dixon, R. A. F., & Darke, P. L. (1990) *J. Biol. Chem. 265*, 14209.

Flaherty, K. M., de Luca-Flaherty, C. R., & McKay, D. B. (1990) *Nature 346*, 623.

Flaherty, K. M., McKay, D. B., Kabsch, W., & Holmes, K. C. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 5041.

Frauenfelder, H., Sligar, S. G., & Wolynes, P. G. (1991) *Science 254*, 1598.

Gerstein, M., & Chothia, C. H. (1991) *J. Mol. Biol. 220*, 133.

Gerstein, M., Schulz, G., & Chothia, C. (1993a) *J. Mol. Biol. 229*, 494.

Gerstein, M., Lesk, A. M., Baker, E. N., Anderson, B., Norris, G., & Chothia, C. (1993b) *J. Mol. Biol. 234*, 357.

Gilliland, G. L., & Quiocho, F. A. (1981) *J. Mol. Biol. 146*, 341.

Harlos, K., Vas, M., & Blake, C. F. (1992) *Proteins: Struct., Funct., Genet. 12*, 133.

Holland, D. R., Tronrud, D. E., Pley, H. W., Flaherty, K. M., Stark, W., Jansonius, J. N., McKay, D. B., & Matthews, B. W. (1992) *Biochemistry 31*, 11310.

Ikura, M., Clore, G. M., Gronenborn, A. M., Zhu, G., Klee, C. B., & Bax, A. (1992) *Science 256*, 632.

Jacobo-Molina, A., Ding, J., Nanni, R. G., Clark, A. D. J., Lu, X., Tantillo, C., Williams, R. L., Kamer, G., Ferris, A. L., Clark, P., Hizi, A., Hughes, S. H., & Arnold, E. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 6320.

Janin, J., & Wodak, S. (1983) *Prog. Biophys. Mol. Biol. 42*, 21.

Joseph, D., Petsko, G. A., & Karplus, M. (1990) *Science 249*, 1425.

Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F., & Holmes, K. C. (1990) *Nature 347*, 37.

Karlsson, R., Zheng, J. H., Xuong, N. H., Taylor, S. S., & Sowadski, J. M. (1993) *Acta Crystallogr. D49*, 381.

Kim, J. L., Nikolov, D. B., & Burley, S. K. (1993) *Nature 365*, 520.

Kim, Y., Geiger, J. H., Hahn, S., & Sigler, P. B. (1993) *Nature 365*, 512.

Kjeldgaard, M., Nissan, P., Thirup, S., & Nyborg, J. (1993) *Structure 1*, 35.

Knighton, D. R., Zheng, J., Ten Eyck, L. F., Ashford, V. A., Xuong, N., Taylor, S. S., & Sowadski, J. M. (1991) *Science 253*, 407.

Knowles, J. R. (1991) *Philos. Trans. R. Soc. London B332*, 115.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A., & Steitz, T. A. (1992) *Science 256*, 1783.

Koshland, D. E., Jr., (1958) *Proc. Natl. Acad. Sci. U.S.A. 44*, 98.

Lawson, C. L., Zhang, R., Schevitz, R. W., Otwinowski, Z., Joachimiak, A., & Sigler, P. B. (1988) *Proteins 3*, 18.

Lebioda, L., & Stec, B. (1991) *Biochemistry 30*, 2817.

Lesk, A. M., & Chothia, C. (1984) *J. Mol. Biol. 174*, 175.

Lesk, A. M., & Chothia, C. (1988) *Nature 335*, 188.

Lindley, P. F., Bajaj, M., Evams, R. W., Garratt, R. C., Hasnain, S., Jhoti, H., Kuser, P., Neu, M., Patel, K., Sarra, R., Strange, R., & Walton, A. (1993) *Acta Crystallogr. D49*, 292.

Loebermann, H., Tokuoka, R., Deisenhofer, J., & Huber, R. (1984) *J. Mol. Biol. 177*, 531.

Lolis, E., & Petsko, G. A. (1990) *Biochemistry 29*, 6619.

Louie, G. V., Brownlie, P. D., Lambert, R., Cooper, J. B., Blundell, T. L., Wood, S. P., Warren, M. J., Woodcock, S. C., & Jordan, P. M. (1992) *Nature 359*, 33.

Luecke, H., & Quiocho, F. A. (1990) *Nature 347*, 402.

Mas, M. T., & Resplandor, Z. E. (1988) *Proteins 4*, 56.

Mas, M. T., Resplandor, Z. E., & Riggs, A. D. (1987) *Biochemistry 26*, 5369.

McPhalen, C. A., Vincent, M. G., Picot, D., Jansonius, J. N., Lesk, A. M., & Chothia, C. (1992) *J. Mol. Biol. 227*, 197.

Meador, W. E., Means, A. R., & Quiocho, F. A. (1992) *Science 257*, 1251.

Meador, W. E., Means, A. R., & Quiocho, F. A. (1993) *Science 262*, 1718.

Milburn, M. V., Tong, L., DeVos, A., Brünger, A., Yamaizumi, Z., Nishimura, S., & Kim, S.-H. (1990) *Science 247*, 939.

Miller, M., Schneider, J., Sathyanarayana, B. K., Toth, M. V., Marshall, G. R., Clawson, L., Selk, L., Kent, S. B. H., & Wlodawer, A. (1989) *Science 246*, 1149.

Montfort, W. R., Perry, K. M., Fauman, E. B., Finermoore, J. S., Maley, G. F., Hardy, L., Maley, F., & Stroud, R. M. (1990) *Biochemistry 29*, 6964.

Mottonen, J., Strand, A., Symersky, J., Sweet, R. M., Danley, D. E., Goeghegan, K. F., Gerard, R. D., & Goldsmith, E. J. (1992) *Nature 355*, 270.

Müller, C. W., & Schulz, G. E. (1988) *J. Mol. Biol. 202*, 909.

Müller, C. W., & Schulz, G. E. (1992) *J. Mol. Biol. 224*, 159.

Noel, J. P., Hamm, H. E., & Sigler, P. B. (1993) *Nature 366*, 654.

Oh, B.-H., Pandit, J., Kang, C.-H., Nikaido, K., Gokcen, S., Ames, G. F.-L., & Kim, S.-H. (1993) *J. Biol. Chem. 268*, 11348.

Olson, A. J., Bricogne, G., & Harrison, S. C. (1983) *J. Mol. Biol. 171*, 61.

Parry, N., Fox, G., Rowlands, D., Brown, F., Fry, E., Acharya, R., Logan, D., & Stuart, D. (1990) *Nature 347*, 569.

Perry, K. M., Fauman, E. B., Finer-Moore, J. S., Montfort, W. R., Maley, G. F., Maley, F., & Stroud, R. M. (1990) *Proteins 8*, 315.

Perutz, M. (1989) *Q. Rev. Biophys. 22*, 139.

Pflugrath, J. W., & Quiocho, F. A. (1988) *J. Mol. Biol. 200*, 163.

Rayment, I., Rypiewski, W. R., Schmidt-Bäse, K., Smith, R., Tomchick, D. R., Benning, M. M., Winkelmann, D. A., Wesenberg, G., & Holden, H. M. (1993) *Science 261*, 50.

Remington, S., Wiegand, G., & Huber, R. (1982) *J. Mol. Biol. 158*, 111.

Rojewska, D., & Elber, R. (1990) *Proteins 7*, 265.

Sack, J. S., Saper, M. A., & Quiocho, F. A. (1989a) *J. Mol. Biol. 206*, 171.

Sack, J. S., Trakhanov, S. D., Tsigannik, I., & Quiocho, F. A. (1989b) *J. Mol. Biol. 206*, 193.

Sali, A., Veerapandian, B., Cooper, J. B., Foundling, S. I., Hoover, D. J., & Blundell, T. L. (1989) *EMBO J. 8*, 2179.

Sali, A., Veerapandian, B., Cooper, J. B., Moss, D. S., Hofmann, T., & Blundell, T. L. (1992) *Proteins: Struct., Funct., Genet. 12*, 158.

Sarra, R., Garratt, R., Gorinsky, B., Jhoti, H., & Lindlay, P. (1990) *Acta Crystallogr. B46*, 763.

Schlichting, I., Almo, S. C., Rapp, G., Wilson, K., Petratos, K., Lentfer, A., Wittinghofer, A., Kabsch, W., Pai, E. F., Petsko, G. A., & Goody, R. S. (1990) *Nature 345*, 309.

Schulz, G. E., Elzinga, M., Marx, F., & Schirmer, R. H. (1974) *Nature 250*, 120.

Schulz, G. E., Muller, C. W., & Diederichs, K. (1990) *J. Mol. Biol. 213*, 627.

Sharff, A. J., Rodseth, L., Spurlino, J. C., & Quiocho, F. A. (1992) *Biochemistry 31*, 10657.

Skarzynski, T., & Wonacott, A. J. (1988) *J. Mol. Biol. 203*, 1097.

Sopkova, J., Renouard, M., & Lewit-Bentley, A. (1993) *J. Mol. Biol. 234*, 816.

Spurlino, J. C., Lu, G. Y., & Quiocho, F. A. (1991) *J. Biol. Chem. 266*, 5202.

Stehle, T., & Schulz, G. E. (1990) *J. Mol. Biol. 211*, 249.

Stein, P., & Chothia, C. (1991) *J. Mol. Biol. 221*, 615.

Stillman, T. J., Baker, B. J., Britton, K. L., & Rice, D. W. (1993) *J. Mol. Biol. 234*, 1131.

Stoddard, B. L., & Koshland, D. E., Jr. (1993) *Biochemistry 32*, 9317.

Taylor, D. A., Sack, J. S., Maune, J. F., Beckingham, K., & Quiocho, F. A. (1991) *J. Biol. Chem. 266*, 21375.

Thayer, M. M., Flaherty, K. M., & McKay, D. B. (1991) *J. Biol. Chem. 266*, 2864.

Vyas, N. K., Vyas, M. N., & Quiocho, F. A. (1988) *Science 242*, 1290.

Vyas, N. K., Vyas, M. N., & Quiocho, F. A. (1991) *J. Biol. Chem. 266*, 5226.

Wade, R. C., Davis, M. E., Luty, B. A., Madura, J. D., & McCammon, J. A. (1993) *Biophys. J. 64*, 9.

Wade, R. C., Brock, A. L., Demchuk, E., Madura, J. D., Davis, M. E., Briggs, J. M., & McCammon, J. A. (1994) *Nature Struct. Biol. 1*, 65.

Weber, I. T., & Steitz, T. A. (1987) *J. Mol. Biol. 198*, 311.

White, J., Hackert, M. L., Buehner, M., Adams, M. J., Ford, G. C., Lentz, P. J. J., Smiley, I. E., Steindel, S. J., & Rossman, M. G. (1976) *J. Mol. Biol 102*, 759.

Wierenga, R. K., Noble, M. E. M., Postma, J. P. M., Groendijk, H., Kalk, K. H., Hol, W. G. J., & Opperdoes, F. R. (1991) *Proteins 10*, 93.

Wodak, S. J., & Janin, J. (1981) *Biochemistry 20*, 6544.

Zhang, R. G., Joachimiak, A., Lawson, C. L., Schevitz, R. W., Otwinowski, Z., & Sigler, P. B. (1987) *Nature 327*, 591.

Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 7558–7562, August 1993
Biophysics

Improved prediction of protein secondary structure by use of sequence profiles and neural networks

(protein structure prediction/multiple sequence alignment)

BURKHARD ROST AND CHRIS SANDER

Protein Design Group, European Molecular Biology Laboratory, D-6900 Heidelberg, Germany

Communicated by Harold A. Scheraga, April 5, 1993

ABSTRACT **The explosive accumulation of protein sequences in the wake of large-scale sequencing projects is in stark contrast to the much slower experimental determination of protein structures. Improved methods of structure prediction from the gene sequence alone are therefore needed. Here, we report a substantial increase in both the accuracy and quality of secondary-structure predictions, using a neural-network algorithm. The main improvements come from the use of multiple sequence alignments (better overall accuracy), from "balanced training" (better prediction of β-strands), and from "structure context training" (better prediction of helix and strand lengths). This method, cross-validated on seven different test sets purged of sequence similarity to learning sets, achieves a three-state prediction accuracy of 69.7%, significantly better than previous methods. In addition, the predicted structures have a more realistic distribution of helix and strand segments. The predictions may be suitable for use in practice as a first estimate of the structural type of newly sequenced proteins.**

The problem of protein secondary-structure prediction by classical methods is usually set up in terms of the three structural states, α-helix, β-strand, and loop, assigned to each amino acid residue. Statistical and neural-network methods use a reduction of the data base of three-dimensional protein structures to a string of secondary-structure assignments. From this data base the rules of prediction are derived and then applied to a test set. For about the last 10 yr, three-state accuracy of good methods has hovered near 62–63%. Recently, values of 65–66% have been reported (1–4). However, when test sets contain proteins homologous to the learning set or when test results have not been multiply cross-validated, actual performance may be lower.

Point of Reference

We use as a "reference network" a straightforward neural-network architecture (5) trained and tested on a data base of 130 representative protein chains (6) of known structure, in which no two sequences have >25% identical residues. The three-state accuracy of this network, defined as the percentage of correctly predicted residues, is 61.7%. This value is lower than results obtained with similar networks (5, 7–10) for the following reasons. (*i*) Exclusion of homologous proteins is more stringent in our data base—i.e., test proteins may not have >30% identical residues to any protein in the training set. Other groups allow cross-homologies up to 49% [e.g., 2-hydroxyethylthiopapain (1ppd) and actinidin (2act) in the testing set termed "without homology" in ref. 5] or 46% (4). (*ii*) Accuracy was averaged over independent trials with seven distinct partitions of the 130 chains into learning and test set (7-fold cross-validation). The use of multiple cross-validation is an important technical detail in assessing performance, as accuracy can vary considerably, depending upon which set of proteins is chosen as the test set. For example, Salzberg and Cost (3) point out that the accuracy of 71.0% for the initial choice of test set drops to 65.1% "sustained" performance when multiple cross-validation is applied—i.e., when the results are averaged over several different test sets. We suggest the term sustained performance for results that have been multiply cross-validated. The importance of multiple cross-validation is underscored by the difference in accuracy of up to six percentage points between two test sets for the reference network (58.3–63.8%).

Use of Multiple Sequence Alignments

It is well known that homologous proteins have the same three-dimensional fold and approximately equal secondary structures down to a level of 25–30% identical residues (11). With appropriate cutoffs applied in a multiple sequence alignment (12), all structurally similar proteins can be grouped into a family, and the approximate structure of the family can be predicted, exploiting the fact that the multiple sequence alignment contains more information about the structure than a single sequence. The additional information comes from the fact that the pattern of residue substitutions reflects the family's protein fold. For example, substitution of a hydrophobic residue in the protein interior by a charged residue would tend to destabilize the structure. This effect has been exploited in model building by homology—e.g. in ref. 13—and in previous attempts to improve secondary-structure prediction (14–18). Our idea was to use multiple sequence alignments rather than single sequences as input to a neural network (Fig. 1). At the training stage, a data base of protein families aligned to proteins of known structure is used (Fig. 2). At the prediction stage, the data base of sequences is scanned for all homologues of the protein to be predicted, and the family profile of amino acid frequencies at each alignment position is fed into the network. The result is striking. On average, the sustained prediction accuracy increases by 6 percentage points. If single sequences rather than profiles are fed into a network trained on profiles, the advantage is generally lost.

Balanced Training

Most secondary-structure prediction methods have been optimized exclusively to yield a high overall accuracy. This method can lead to severe artifacts because of the very uneven distribution of secondary-structure types in globular proteins: 32% α-helix, 21% β-strand, and 47% loop (our data base). Usually, loops are predicted quite well, helices are predicted medium well, and strands are predicted rather poorly. This imbalance can be corrected if the network is

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

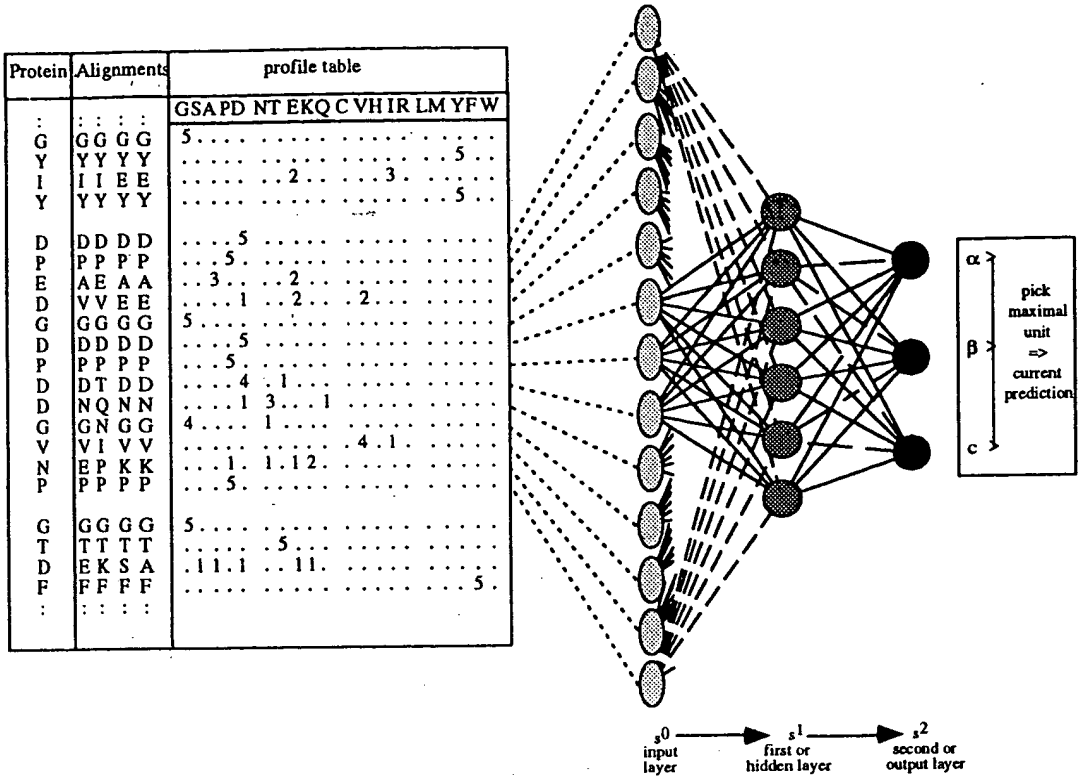

FIG. 1. Network architecture. A sequence profile of a protein family, rather than just a single sequence, is used as input to a neural network for structure prediction. Each sequence position is represented by the amino acid-residue frequencies derived from multiple sequence alignments as taken from the homology-derived structure of proteins (HSSP) data base (12). The residue frequencies for the 20-residue types are represented by 3 bits each (or by one real number). To code the N- and C-terminal ends an additional 3 bits are required (or one real number). The 63 bits originating from one sequence position are mapped onto 63 (21 for real numbers) input units of the neural network. A window of 13 sequence positions, thus, corresponds to 819 (273) input units. The input signal is propagated through a network with one input, one hidden, and one output layer. The output layer has three units corresponding to the three secondary-structure states, helix, β-strand, and "loop," at the central position of the input sequence window. Output values are between 0 and 1. The experimentally observed secondary structure states (19) are encoded as 1,0,0 for helix; 0,1,0 for strand; and 0,0,1 for loop. The error function to be minimized in training is the sum over the squared difference between current output and target output values. Net cascade: the first network (sequence-to-structure) is followed by a second network (structure-to-structure) to learn structural context (not shown). Input to the second network is the three output real numbers for helix, strand, and loop from the first network, plus a fourth spacer unit, for each position in a 17-residue window. From the $17 \times (3 + 1) = 68$ input nodes the signal is propagated via a hidden layer to three output nodes for helix, strand, and loop, as in the first network. In prediction mode, a 13-residue sequence window is presented to the network, and the secondary-structure state of the central residue is chosen, according to the output unit with the largest signal.

trained with each type of secondary structure in equal proportion (33%), rather than in the proportion present in the data base or anticipated in the proteins to be predicted. The result is a more balanced prediction (Fig. 3; Table 1), without affecting, negatively or positively, the overall three-state accuracy. A similar result was reported by Hayward and Collins (22). The main improvement is in a better β-strand prediction, the most difficult of the three states to predict. The method maintains full generality—i.e., it is equally applicable to all-α, mixed $\alpha\beta$, and all-β proteins. No knowledge of the structural type of the protein is required, as is the case for methods optimized on particular structural classes (9, 23).

Training on Structural Context

Even if a prediction method has high overall accuracy and is well balanced, it can be woefully inadequate in the length distribution of the predicted helices and strands. For example, the reference network predicts too many short strands and helices and too few long ones (Fig. 4). The predictions of this network appear fragmented compared with typical globular proteins. Published prediction methods have similar shortcomings in the length distribution of segments to various extents, except for two methods that optimize the sum of segment scores by dynamic programming (W. Kabsch and C.S., personal communication and ref. 24). The shortcoming is partly overcome here by feeding the three-state prediction output of the first, "sequence-to-structure," network, into a second, "structure-to-structure," network. The second network is trained to recognize the structural context of single-residue states, without reference to sequence information. Training it is very similar to that used for the sequence-to-structure network. The output string of the first network—e.g., the partially incorrect string HHHEEHH (two β-strand residues in the middle of a helix)—becomes the input to the second network and is confronted with correct structure HHHHHHH, a helical segment. Network couplings are optimized to minimize the discrepancy. The addition of the structure–structure network increases the overall accuracy only marginally but reproduces substantially better the length distribution of helices and strands. A simple way of measuring the quality of segment lengths is to compare the average length of helices and strands in the data base to those in the predicted structures ($\langle L_\alpha \rangle = 6.9$, $\langle L_\beta \rangle = 4.6$, Fig. 4). A similar second-level network was used by Qian and Sejnowski (5), but no effect of improved prediction of segment lengths was reported.

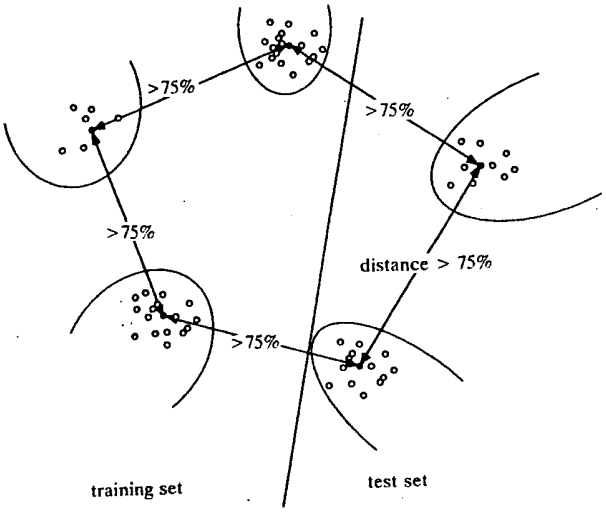

FIG. 2. Partition of protein families into training and test set. The structurally known representatives of the families used for training the network have a distance of at least 75% to those used for testing (sequence distance in percent nonidentical residues; drawn schematically). Each family contains homologous sequences, defined as those with a sequence identity >30% to the representative. 3D, three dimensional.

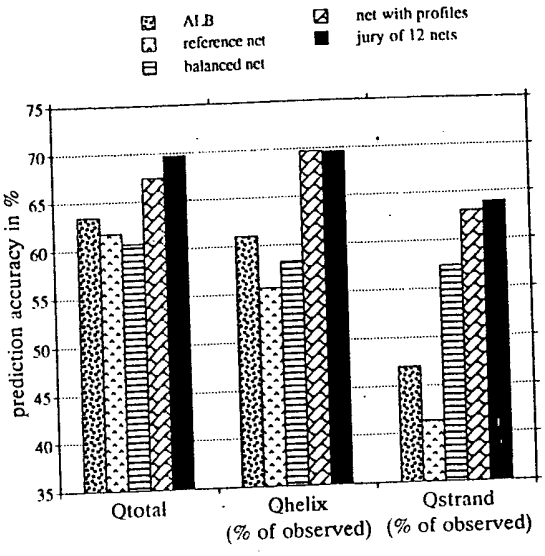

FIG. 3. Testing five secondary-structure prediction methods on the same set of proteins reveals the contribution of different devices to the improvement of accuracy. Q_{total}, overall prediction accuracy for the three states (helix, strand, loop; number of residues predicted correctly divided by the total number of residues). Q_{helix} and Q_{strand}, prediction accuracy calculated separately for helix and strand (e.g., number of helix residues predicted correctly divided by number of observed helix residues). The methods tested on our data base are ALB (20), first-level network with no balanced learning and no profiles (reference net), a two-level network cascade with balanced learning and no profiles (balanced net), a two-level network cascade with profiles and balanced learning (net with profiles), and 12 different networks combined by majority vote (jury of 12 nets). Some groups achieve higher accuracy than does ALB, but the accuracy values are not strictly comparable, as they are based on different test data sets and, in part, on test proteins with detectable sequence similarities to proteins on which the method was trained. Values for Q_{total} (Q_{helix}, Q_{strand}, Q_{loop}) are 65.5% (65, 45, 74), COMBINE (2); 63.0% (58, 54, 68), SIMPA (1); and 66.4%, Zhang *et al.* (4). Observed versus predicted matrix for the best method is indicated in Table 1.

"Jury of Networks"

An additional two percentage points in overall accuracy were gained by a jury of networks that predicts by simple majority vote of a set of 12 different networks. The increased accuracy is an effect of noise reduction, mitigating the ill effects of incomplete optimization when any single network settles into a local minimum of the error function.

Overall Improvement

The final jury of networks outperforms all known methods in overall accuracy, balanced β-strand prediction, and length distribution of segments as follows.

(*i*) The overall accuracy is 69.7%, three percentage points above the highest value reported so far [66.4% (4)]. The actual improvement may be larger, as their test set has sequence similarities of up to 46% relative to the training set. The improvement is six percentage points relative to the best classical method tested on our data base [63.4%, ALB (20)]. For a new protein sequence, one can expect a prediction accuracy between 61% and 79% (1 SD about the average over individual proteins of 70.2%), provided several homologous sequences are available. Values for three-state accuracy should not be confused with those for two-state accuracy (9, 23). Two-state predictions—e.g., for the state helix/nonhelix—carry less information and have a base value for random prediction of 50%—i.e., 17 percentage points higher than that for three-state methods.

(*ii*) Accuracy is well-balanced at 70% helix and 64% strand, measured as the percentage "correct of *observed*" (Fig. 3). The percentages "correct of predicted"—i.e., the probability of correct prediction, given a residue predicted in a particular state—are 72% helix and 57% strand.

(*iii*) The length distribution of segments is more "protein-like" (Fig. 4). Unfortunately, the length distribution is not

Table 1. Observed versus predicted matrix for best method of Fig. 3

	Residues predicted			Total observed	Residues predicted correctly,* %	
	Helix	Strand	Loop		Of observed	Of predicted
Residues observed						
Helix	5552	774	1,646	7,972	70	72
Strand	517	3229	1,310	5,056	64	57
Loop	1548	1592	8,227	11,367	72	73
Total predicted	7617	5595	11,183	24,395		
Correlation coefficient (21)	0.58	0.50	0.50			

*Prediction of jury of 12 nets method.

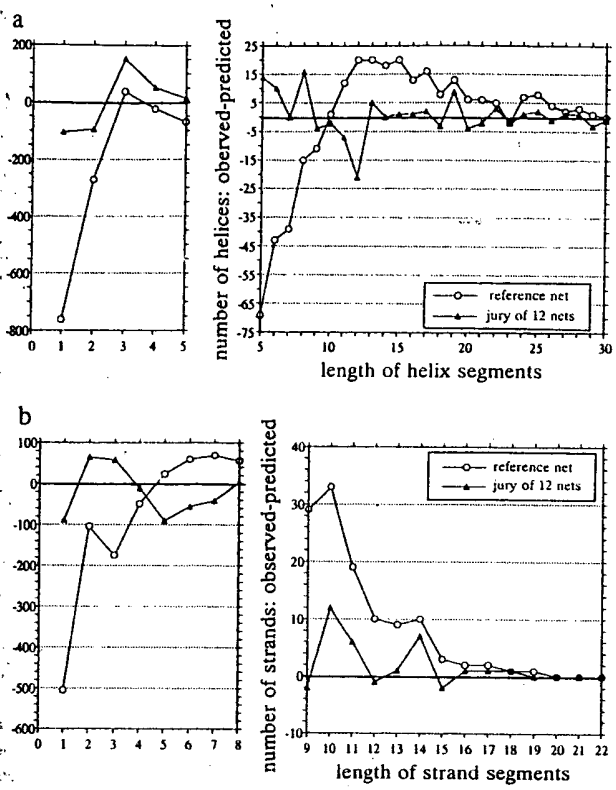

FIG. 4. Deviation in the length distribution of observed and predicted segments is an additional criterion by which prediction methods can be evaluated. (*a*) Difference in the length distribution of helix segments—i.e., number of observed segments in a given length range minus number of predicted segments. (*b*) Difference in the length distribution of strand segments. Predictions by the simple net (no profile, not balanced, no cascade) result in too many short segments, too few long segments; prediction by the jury of 12 nets results in a length distribution much closer to the observed one. Average segment lengths are as follows: reference net, $\langle L_\alpha \rangle = 4.2$ and $\langle L_\beta \rangle = 2.9$ residues; jury of 12, $\langle L_\alpha \rangle = 8.9$ and $\langle L_\beta \rangle = 5.1$ (observed: $\langle L_\alpha \rangle = 9.0$ and $\langle L_\beta \rangle = 5.1$).

generally given in the literature, but most methods are inferior in this regard.

Tests on Completely New Proteins

How accurate are predictions likely to be in practice? As a final check, the network system was trained on the full set of 151 sequence families of known structure and then tested on 26 protein families for which a first x-ray or NMR three-dimensional structure became available after the network architecture had been finalized. None of these additional test proteins had >25% sequence identity relative to any of the training proteins (Fig. 5). In this final set, 72% of the observed helical and 68% of strand residues were predicted correctly. The overall three-state accuracy for this set of completely new protein structures was 70.3%.

Predictions via Electronic Mail

Secondary-structure predictions using the currently best version of the profile network from Heidelberg (PHD) are available via electronic mail. Send a message containing the word "help" to PredictProtein@EMBL-Heidelberg.de. In practice, the predictions give a good first hypothesis of the structural properties of any newly sequenced water-soluble protein and may be an aid in the planning of point-mutation experiments and in the prediction of tertiary structure.

Conclusion

There are two important practical limitations: most of the advantage of the current method is lost when no sequence homologues are available; and the method in its current implementation is not valid for membrane proteins and other nonglobular or non-water-soluble proteins.

A major limitation in principle of the current method lies in its limited goal: secondary structure is a very reduced description of the complexities of three-dimensional structure and carries little information about protein function. However, as long as reliable prediction methods for protein three-dimensional structure and function are not available, secondary-structure predictions of improved quality are useful in practice—e.g., for the planning of point-mutation experiments, for the selection of antigenic peptides, or for identification of the structural class of a protein. Indeed, interest in the community is substantial: during 6 mo. since submission of this manuscript, >3,000 predictions for a wide variety of sequences have been requested and served via electronic mail.

Looking ahead, we would not be surprised to see increasingly successful use of evolutionary information in attempts to predict more complex aspects of protein structure and function. Sequence families grouped around one structure as well as structural superfamilies with common folds but divergent sequences (26, 27) contain a wealth of information not available 14 yr ago at the time of the first attempts at using homologous sequences for improved prediction (16). Having posed the puzzle of protein folding, evolution may hand us the key to its successful solution.

Note Added in Proof. Since the submission of this paper (April 1993) the method described has been improved further. By explicitly using

```
number      .........1.........2.........3.........4.........5.........6.........7.........8
sequence    AFDGTWKVDRNENYEKFMEKMGINVVKRKLGAHDNLKLTITQEGNKFTVKESSNFRNIDVVFELGVDFAYSLADGTELTG
observed       EEEEEEEEE HHHHHHH     HHHHHHH     EEEEEEE  EEEEEEE    EEEEEEE    EEEE      EEEE
predicted         EEEE   HHHHHHHHHHHHHHHHHHHHH   EEEEEE  EEEEEEE  EEEEEEEEEE  EHHEE        EE

number      .........9.........0.........1.........2.........3.
sequence    TWTMEGNKLVGKFKRVDNGKELIAVREISGNELIQTYTYEGVEAKRIFKKE
observed    EEEEE  EEEEEEEE      EEEEEEEEE  EEEEEEEE  EEEEEEEEE
predicted   EEEE   HHEEEEEE      HHHHHHHHH    EEEEEE EEEEEE
```

FIG. 5. Example of prediction for a protein sequence by the currently best method. The β-barrel structure of intestinal fatty acid-binding protein has just become available through Protein Data Bank [code 1ifb (25)]. Prediction accuracy is 71.8%. In this β-sandwich structure, 8 out of the 10 β-strands are predicted correctly (one strand is ambiguous, and one strand is predicted as helix, but the ends of the segment are correct), and the two helices are predicted as one long helix (E: strand, H: helix). For all 26 new protein chains, including 1ifb, overall accuracy averaged over single residues is 70.3%; averaged over single proteins, it is 71.1%. The estimated probabilities of correct prediction, given a residue *predicted* in a helix, strand, or loop were 69%, 58%, or 77%, respectively (see text for probabilities relative to the number of residues *observed* in the three states). These 26 protein chains were not available publicly at the time of development of the method and were only used once in a final test of the currently best method. They are as follows: 1ace, 1cox, 1cpk_E, 1dfn_B, 5enl, 1f3g, 3fgf, 2gbl, 1gly, 1gmf_A, 1hcc, 1hdd_C, 2hip_B, 1ifb, 1msb_A, 1nsb_B, 5p21, 1pi2, 2pk4, 1rop_A, 1sar_A, 2scp_A, 1snv, 3trx, 3znf, 2zta_A (all taken from the Protein Data Bank prerelease of July 1992; membrane proteins and proteins with many metals or SS bridges were not considered).

conservation weights and the numbers of insertions and deletions in the multiple sequence alignments as input to the network system, the sustained overall three-state accuracy becomes 71.4% on the same data set used in this paper.

We thank Gerrit Vriend and Reinhard Schneider for stressing the importance of sequence profiles and segment lengths and Michael Scharf for general support; L. Philipson for reducing administrative load; and the Human Frontiers Science Program and the European Community Bridge Program for financial support.

1. Garnier, J. & Levin, J. M. (1991) *Comput. Appl. Biosci.* **7**, 133–142.
2. Levin, J. M. & Garnier, J. (1988) *Biochim. Biophys. Acta* **955**, 283–295.
3. Salzberg, S. & Cost, S. (1992) *J. Mol. Biol.* **227**, 371–374.
4. Zhang, X., Mesirov, J. P. & Waltz, D. L. (1992) *J. Mol. Biol.* **225**, 1049–1063.
5. Qian, N. & Sejnowski, T. J. (1988) *J. Mol. Biol.* **202**, 865–884.
6. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992) *Protein Sci.* **1**, 409–417.
7. Holley, H. L. & Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 152–156.
8. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Lautrup, B., Nørskov, L., Olsen, O. H. & Petersen, S. B. (1988) *FEBS Lett.* **241**, 223–228.
9. Kneller, D. G., Cohen, F. E. & Langridge, R. (1990) *J. Mol. Biol.* **214**, 171–182.
10. Stolorz, P., Lapedes, A. & Xia, Y. (1992) *J. Mol. Biol.* **225**, 363–377.
11. Chothia, C. & Lesk, A. M. (1986) *EMBO J.* **5**, 823–826.
12. Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56–68.
13. Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L. (1990) *Proc. R. Soc. London B* **241**, 132–145.
14. Benner, S. A. & Gerloff, D. (1990) *Adv. Enzyme Regul.* **31**, 121–181.
15. Barton, G. J., Newman, R. H., Freemont, P. S. & Crumpton, M. J. (1991) *Eur. J. Biochem.* **198**, 749–760.
16. Maxfield, F. R. & Scheraga, H. A. (1979) *Biochemistry* **18**, 697–704.
17. Russell, R. B., Breed, J. & Barton, G. J. (1992) *FEBS Lett.* **304**, 15–20.
18. Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987) *J. Mol. Biol.* **195**, 957–961.
19. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22**, 2577–2637.
20. Ptitsyn, O. B. & Finkelstein, A. V. (1983) *Biopolymers* **22**, 15–25.
21. Matthews, B. W. (1975) *Biochim. Biophys. Acta* **405**, 442–451.
22. Hayward, S. & Collins, J. F. (1992) *Proteins* **14**, 372–381.
23. Muggleton, S., King, R. D. & Sternberg, M. J. E. (1992) *Protein Eng.* **5**, 647–657.
24. Schneider, R. (1989) Diploma thesis (Dept. of Biology, Univ. Heidelberg, F.R.G.).
25. Sacchettini, J. C., Gordon, J. I. & Banaszak, L. J. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 7736–7740.
26. Richardson, J. (1981) *Adv. Protein Chem.* **34**, 168–339.
27. Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992) *Protein Sci.* **1**, 1691–1698.

PROTEINS: Structure, Function, and Genetics 19:141–149 (1994)

Accuracy of Protein Flexibility Predictions

Mauno Vihinen,[1] Esa Torkkila,[2] and Pentti Riikonen[2]
[1]Department of Biochemistry, University of Turku, SF-20500 Turku and Turku Centre for Biotechnology, University of Turku, SF-20520 Turku and [2]Department of Computer Science, University of Turku, SF-20520 Turku, Finland

ABSTRACT **Protein structural flexibility is important for catalysis, binding, and allostery. Flexibility has been predicted from amino acid sequence with a sliding window averaging technique and applied primarily to epitope search. New prediction parameters were derived from 92 refined protein structures in an unbiased selection of the Protein Data Bank by developing further the method of Karplus and Schulz (Naturwissenschaften 72:212–213, 1985). The accuracy of four flexibility prediction techniques was studied by comparing atomic temperature factors of known three-dimensional protein structures to predictions by using correlation coefficients. The size of the prediction window was optimized for each method. Predictions made with our new parameters, using an optimized window size of 9 residues in the prediction window, were giving the best results. The difference from another previously used technique was small, whereas two other methods were much poorer. Applicability of the predictions was also tested by searching for known epitopes from amino acid sequences. The best techniques predicted correctly 20 of 31 continuous epitopes in seven proteins. Flexibility parameters have previously been used for calculating protein average flexibility indices which are inversely correlated to protein stability. Indices with the new parameters showed better correlation to protein stability than those used previously; furthermore they had relationship even when the old parameters failed.**

Key words: dynamics, flexibility index, protein stability, antigenic regions, epitopes

INTRODUCTION

Protein molecules are dynamic being in constant motion. Structural flexibility is essential for activity but, on the other hand, structural stability requires rigidity.[1–3] Flexible regions are found in catalytic sites,[4–7] binding sites,[8] antigenic regions,[9] sites susceptible for proteolytic cleavage,[10] allosteric hinge sites,[11] etc. Proteins with similar functions have similar excess of flexibility in their optimum reaction conditions.[2,4]

The core of a globular protein is relatively tightly packed. Surface residues are generally more mobile due to fewer stabilizing interactions. Exposed surface loops are the most flexible and show the largest sequence variation. The time scale of protein mobility is very wide, the fastest vibrations and motions requiring only 10^{-14} to 10^{-13} s. Mobility can be simulated with molecular dynamics. Although the simulations are relatively short plenty of valuable information is available. The flexible regions can be predicted using less accurate methods even without structural information.

Three techniques have been used for predicting protein flexibility from amino acid sequence. The methods of Karplus and Schulz[12] (KS) and of Bhaskaran and Ponnuswamy[13] (BP) are based on parameters derived from three-dimensional structures. Ragone et al.[14] (R) base their approach on a combination of hydropathy predictions and amino acid volumes. Flexibility analysis can be used to search for the most mobile and thus possibly also the surface residues in a sequence, which are thought to represent epitopes. For vaccine production, it would be of great value to be able to predict the antigenic regions of a protein from its sequence. Flexibility predictions have been used in searching for continuous epitopes from amino acid sequences.[12,14,15] Other epitope prediction methods include hydropathy,[16,17] β-turn propensity,[18] and joint prediction of hydropathy, surface accessibility, flexibility, and secondary structure.[19] Stern[20] has recently reviewed the methods.

The KS method uses normalized *B*-values of C_α-atoms in 31 protein structures. Here we have extended the flexibility prediction by analyzing all the backbone atoms of 92 well refined structures from an unbiased selection of PDB.[21] To test the applicability of the predictions they were compared to ex-

Abbreviations: BP, flexibility according to Bhaskaran and Ponnuswamy; KS, according to Karplus and Schulz; R, according to Ragone et al.; VTR, according to parameters derived here; PDB, Protein Data Bank.

Received October 5, 1993; revision accepted February 1, 1994.

Address reprint requests to Mauno Vihinen at his present address: Center for Structural Biochemistry, Karolinska Institute, Novum, S-14157 Huddinge, Sweden.

perimental B-values. Their use in predicting antigenic regions was studied with proteins for which locations of continuous epitopes have been determined. Increased hydrophobicity and decreased flexibility have been shown to be the main stabilizing principles in thermostable proteins.[22] Previously we have shown inverse correlation between thermal stability and structural flexibility by calculating flexibility indices from normalized B-values for amino acid sequences.[2] Even a better and more accurate correlation to stability was noticed when flexibility indices were calculated with the new parameters.

METHODS

Entries for high resolution structures containing B-values were taken from the unbiased selection of Protein Data Bank[21] (PDB) because there are a lot of redundant data in PDB. Only 78 of the original 102 entries could be used in this analysis because of missing or incomplete B-values or sequence information. Some of the chosen entries contained several proteins, so finally there were 92 different structures. The PDB entries were 1bp2, 1ccr, 1cla, 1cse 2 chains), 1ctf, 1eca, 1fc2 (2 chains), 1gcr, 1gd1, 1gox 1gp1, 1hoe, 1i1b, 1ldm, 1lz1, 1mbd, 1nxb, 1pcy, 1phh, 1prc (4 chains), 1r69, 1sgt, 1sn3, 1tnf, 1ubq, 1utg 1wsy (2 chains), 2aza, 2cab, 2ccy, 2cdv, 2ci2, 2cpp 2cts, 2cyp, 2fb4 (2 chains), 2gbp, 2gn5, 2hhb (2 chains), 2hla (2 chains), 2hmg (2 chains), 2lbp, 2lh2, 2ltn 2 chains), 2mhr, 2ovo, 2paz, 2pfk, 2rnt, 2rsp, 2sga, 2sod, 2ts1, 2wrp, 3adk, 3gap, 3grs, 3ins (2 chains), 3lzm, 3rn3, 3tln, 451c, 4cha, 4fd1, 4hvp, 4pep, 4xia, 5at1 (2 chains), 5cpa, 5pti, 5rxn, 6acn, 7api 2 chains), 8adh, 8cat, 8dfr, 9pap, 9wga. Normalized B-values derived from the unbiased structures were used both for flexibility prediction and the calculation of flexibility indices. Computer programs were developed to be compatible with the GCG program suite.[23]

Calculation of Normalized B-Values

The selection of 92 unbiased protein structures was used to derive normalized B-values. Temperature factors of the backbone atoms N, C_α, C, and O were taken from the PDB.[24] The Karplus and Schulz[12] approach of determining normalized B-values was repeated with our extended database. The threshold values are those previously used. The B-values of each protein were normalized so that the mean was 1.0 and the root mean square deviation 0.3. Based on its deviation from the mean, each residue type was defined as flexible or rigid. Those with average B_{norm} values below 1.0 were denoted as rigid. In the next step normalized B-values were determined for each residue type when surrounded by none, one or two rigid neighbours to obtain B_{norm0}, B_{norm1}, and B_{norm2} tables, respectively. Because chain termini are usually very flexible and could have caused bias, three N- and C-terminal residues were omitted from each structure.

Programs for Flexibility Prediction

Program FLEX was implemented for flexibility predictions with our new B_{norm} tables and with the parameters of Bhaskaran and Ponnuswamy[13] and Ragone et al.[14] The antigenic index of Jameson and Wolf[19] was also included. The sequence can be read either from a PDB or a GCG file. The predictions are based on a sliding window averaging technique. The optimized window size for each technique is used: five for R, seven for BP, and nine residues for our parameters and those of KS. The propensities for the residues inside the window are summed up and given for the residue in the middle of the window. The weighting of residues inside the window is 0.25, 0.4375, 0.625, 0.8125, 1, 0.8125, 0.625, 0.4375, and 0.25 from left to right in techniques using B_{norm} values but has a constant value of 1 in the methods of BP and R. The flexibilities of KS and ours are calculated as follows. First the number of rigid neighbors around each residue is determined. Then the neighbor correlated weighted propensities from B_{norm} tables are summed and given for the middlemost residue after which the window is shifted by one residue. The results can be presented with the program FLEXPLOT on several graphics devices. Experimental B-values are shown for the backbone atoms of proteins in PDB entries.

Testing Accuracy of the Flexibility Predictions

The accuracy of the different flexibility prediction methods was studied by determining correlation coefficients. The B-values for each of the proteins were compared to predicted flexibilities by calculating correlation coefficients. Many PDB structures contain one or just few highly flexible residues due to, e.g., lattice disorders. To see if the high peaks might bias the analysis, the B-values of residues in each protein were scaled from 0 to 100%. If only one residue had flexibility higher than 80 or 90%, its value was reduced to that of the second highest residue and the analysis was repeated until there were residues also on intervals 80 to 90% and/or 90 to 100%. The correlation coefficients were determined for the entries both when smoothed by sieving the high peaks and when untreated.

Optimization of the Flexibility Prediction Techniques

The flexibility prediction techniques use the sliding window averaging technique. The only adjustable parameter in the R and BP methods is the width of the window, i.e., number of consecutive residues used in the prediction at a time. In the method of KS the window was originally fixed to seven residues but the residues inside the window had differ-

ent weighting depending on their location within the window. The length of the window was optimized for all four prediction techniques by determining correlation coefficients and maximizing information contents with window lengths 5 to 15 residues. In addition also the effect of residue weighting was studied by giving the weight of 0.25 for the first and last residues in the window and 1.0 for the middlemost. The weights of the others were at equal spacing between these two values.

Calculation of Flexibility Indices

Atomic temperature factors (*B*-values) obtained during crystal structure determination are a measure of the flexibility of the residues in the protein. We have used normalized *B*-values to calculate average flexibility indices for the whole protein molecule. Since the flexibility of a residue is dependent on the nature of neighboring residues, three parameter tables are used. There have been two ways to calculate average flexibility indices.[2] The *F* index is calculated from

$$F = \sum_{i=2}^{n-1} B_{\mathrm{nc}\,i}/(n-2)$$

where n is the number of residue and B_{nc} is neighbor correlated normalized *B*-value for the residue type. Another equation, F_7, gives different emphasis for the chain termini

$$F_7 = \sum_{i=8}^{n-7} f_i/(n-8)$$

where $f_i = [B_{\mathrm{nc},i} + 0.75(B_{\mathrm{nc},i-1} + B_{\mathrm{nc},i+1}) + 0.5(B_{\mathrm{nc},i-2} + B_{\mathrm{nc},i+2}) + 0.25(B_{\mathrm{nc},i-3} + B_{\mathrm{nc},i+3})]/4$. Now that window size nine was found to be optimal in predictions with normalized *B*-values a new equation was determined

$$F_9 = \sum_{i=10}^{n-9} f_i/(n-10)$$

where $f_i = [B_{\mathrm{nc},i} + 0.8125(B_{\mathrm{nc},i-1} + B_{\mathrm{nc},i+1}) + 0.625(B_{\mathrm{nc},i-2} + B_{\mathrm{nc},i+2}) + 0.4375(B_{\mathrm{nc},i-3} + B_{\mathrm{nc},i+3}) + 0.25(B_{\mathrm{nc},i-4} + B_{\mathrm{nc},i+4})]/5.25$.

RESULTS AND DISCUSSION

New Flexibility Parameters

Three methods have been used to predict protein structural flexibility from sequences.[12–14] The parameters for the KS- and BP-techniques were derived from known 3D structures, whereas those for the R-technique are combined from other predictions. A limited set of 31 structures was used in the KS method to determine prediction parameters, whereas BP had only 19 proteins. We have extended the analysis to 92 refined structures. We reimplemented the KS algorithm because we found it gave the most accurate predictions. All these techniques use a sliding window averaging technique; parameters are summed for a stretch of amino acids within a window which is shifted by one residue at a time. In the KS method residues have coefficients dependent on the location within the window, thus the contribution of a residue to the prediction value depends on its distance from the middle of the window. Claverie and Daulmerie[25] argue that smoothing of the prediction curves by weighting is advantageous, since pattern recognition is easier and the irregular variation of values is damped. The smoothing is also better for detecting local maxima which are of importance, e.g., in epitope analysis. von Heijne[26] has used a related trapezoid weighting scheme in analysis of membrane spanning segments.

The proteins for calculating normalized *B*-values were taken from an unbiased selection of the PDB.[21] Normalized *B*-values, hereafter VTR parameters, were calculated from the 92 structures (Table I). The major difference to those of KS is that we have 11 rigid residues instead of 10. Threonine is classified as a rigid amino acid, because its average B_{norm} is below 1. The order and values of residues have changed. Glycine is generally considered to be the most flexible amino acid. It has the highest value both in BP and R tables but not in the KS table. In our analysis it is found to be flexible but there are still seven more flexible residue types. This might be because the more flexible residues, which are all charged or polar except for proline, appear mainly on surface whereas glycine is also found in the protein interior. As the normalized *B*-values are averages the restricted mobilities of buried glycine residues may reduce the overall value. Another explanation might be frequent occurrence in tight turns having restricted mobility. The values for glycine are the most neighbour dependent. When surrounded by one or two rigid residues it is among the most flexible residues.

The new B_{norm} values were used in flexibility prediction. If the sequence in program FLEX is read from PDB file *B*-values are averaged for the backbone atoms. The predictions and the *B*-values are presented with program PLOTFLEX. For the plots the values of BP and R tables were normalized to be from 0 to 1. In the original R parameterization the most flexible residue had the lowest value thus the numbers were inversed to be comparable to the others.

Accuracy of the Flexibility Prediction Techniques

The accuracy of the techniques was tested with correlation coefficients method. The prediction window was adjusted from 5 to 15 residues and the prediction accuracy was followed when the highest *B*-value peaks were either smoothed or not. The means of the correlation coefficients over the 92 proteins in Table II shows that the optimal window in

TABLE I. Neighbor Correlated Normalized Flexibility Parameters of the 92 Protein Structures

Residue	Count	$B_{norm,avr}$	B_{norm0}		B_{norm1}		B_{norm2}	
			Count	Value	Count	Value	Count	Value
W	264	0.904	51	1.186	60	0.938	153	0.796
C	333	0.906	68	1.196	79	0.939	186	0.785
F	708	0.915	159	1.247	154	0.934	395	0.774
I	926	0.927	208	1.241	213	0.977	505	0.776
Y	646	0.929	144	1.199	165	0.981	337	0.788
V	1297	0.931	296	1.235	325	0.968	676	0.781
L	1505	0.935	346	1.234	365	0.982	794	0.783
H	457	0.950	112	1.279	121	0.967	224	0.777
M	349	0.952	85	1.269	75	0.963	189	0.806
A	1499	0.984	432	1.315	338	0.994	729	0.783
T	1057	0.997	300	1.324	270	0.998	487	0.795
R	764	1.008	225	1.310	186	1.026	353	0.807
G	1529	1.031	491	1.382	359	1.018	679	0.784
Q	674	1.037	213	1.342	161	1.041	300	0.817
S	1171	1.046	378	1.381	279	1.025	514	0.811
N	794	1.048	255	1.380	221	1.022	318	0.799
P	857	1.049	295	1.342	201	1.050	361	0.809
D	1011	1.068	371	1.372	226	1.022	414	0.822
E	1027	1.094	396	1.376	253	1.052	378	0.826
K	1038	1.102	420	1.367	278	1.029	340	0.834

BP method was seven residues and five in R. We can see that the R method is the overall poorest technique. This can be understood from the origin of the parameters which were obtained by multiplying residue hydrophobicities by volumes without using any structural analysis. The three methods giving better correlation coefficients are based on three-dimensional structures.

The optimization of the predictions with the VTR and KS parameters required weighting of the residues in the window. The window sizes tested were from 5 to 15 residues. The weights of the outermost residues were 0.25 and 1 for that in the middle. The value 0.25 was chosen to give some emphasis also for the ends of windows. The optimal window size was nine for both VTR and KS techniques (Table II), the latter of which previously used seven residues. The correlation coefficients with the optimized prediction techniques for all the 92 protein structures with all the four techniques are as follows: VTR 0.3304, KS 0.3356, BP 0.2428, and R 0.1659. Clearly the best results were obtained with the VTR and KS methods, the other two being much poorer. The predictive power varies greatly in each technique. The best correlation coefficients are close to 0.8 whereas the poorest values are close to 0. All the tested methods predicted poorly some proteins. The results show that in the KS method the previously used window of 7 residues is not optimal. The best results are obtained with nine consecutive amino acids. The new parameters were better than those of KS with short (5 to 7 residues) and large (15 residues) window but the differences are not significant.

Many proteins have one or only a few residues with very large B-values due to, e.g., static disorders in crystal lattice. These residues produce high peaks on B-value curves and might bias the parameters for those residues. This could be avoided by smoothing the curve, but no real effect on predictability was seen, e.g., in the case of the VTR method the unsmoothed value with window size 9 was 0.3302 while it was 0.3304 for the smoothed data. The same order of improvement was noticed also in the other three methods. Many of the highest peaks were already filtered away when the three N- and C-terminal residues of each protein were not included in the calculation of prediction parameters. This was done because the ends are known to be exceptionally flexible.

The use of backbone atoms was tested comparing predictability to parameters derived from C_α atoms of the 92 proteins. Correlation coefficients were determined for both parameter sets and the mean was found to be 0.330 for the backbone derived data with window 9 whereas is was 0.320 for the tables derived from C_α atoms. The improvement in the backbone-derived data is surprisingly small. The same result can be noticed when comparing the results of backbone parameters (VTR scale) to those of C_α parameters (KS table). It seems that in KS analysis there were enough data to bring the predictability with this sort of technique close to its maximum and our data for 17,906 amino acids did not change it much.

We tested the prediction methods also with structures not included in our data set; 38 randomly selected structures from a later version of PDB not having significant sequence similarity to the proteins used in the derivation of the parameters were

TABLE II. Optimization of Flexibility Prediction Windows*

Window size	Prediction technique			
	VTR	KS	BP	R
5	0.3158	0.3112	0.2345	0.1659
7	0.3266	0.3283	0.2428	0.1655
9	0.3304	0.3356	0.2387	0.1644
11	0.3280	0.3332	0.2219	0.1602
13	0.3204	0.3235	0.2092	0.1628
15	0.3125	0.3142	0.2030	0.1645

*The overall correlation coefficients for all the chosen 92 proteins were determined with different prediction window sizes.

analyzed. Here, too, predictions with the VTR and KS methods are giving best results (mean values 0.3359 and 0.3260, respectively), R and BP scales are clearly the worst ones (mean values 0.2460 and 0.2596, respectively). The accuracy of the predictions are of the same order as for the structures used to derive the parameters, but the differences are not significant. The VTR is somewhat better than the KS method. The most striking result is an increase in the predictability of the R method. VTR and KS parameters ar the best and the new parameters are somewhat more accurate.

The applicability of the flexibility predictions is shown for myoglobin in Figure 1. The VTR and KS plots resemble each other although the new scale is discriminating flexible and rigid regions more sharply, which is advantageous in searching for antigenic regions. The flexibilities of the two techniques follow quite well the shape of the *B*-value curves, although the predicted curves are smoother.

The flexibility predictions and experimental *B*-values could further be compared with the program MULTICOMP,[27] a multiple sequence comparing tool which can also be used for comparing predictions. Prior to this kind of analysis the *B*-values and flexibility propensities have to be normalized to express the same range of values. This approach has also been used to compare hydropathy predictions by comparing two different methods of predicting hydropathic character on the same protein.[28]

Prediction of Antigenic Sites

The protein surface serves as a template for numerous antibodies. Some of the epitopic regions are formed by consecutive residues. These regions have been determined for several proteins such as sperm whale myoglobin[16] (PDB entry 1 mbo), hen egg white lysozyme[16] (1lyz), tobacco mosaic virus protein[29] (2tmv), horse cytochrome *c*[16] (sequence entry ccho), bovine serum albumin[30] (a36401), rotavirus major outer-shell glycoprotein[31] (vs09_rots1), and hepatitis B virus core protein[32] (nkvlah). The proteins contained although 31 continuous epitopes when the N- and C-terminal regions were omitted. Since one or the major applications of the flexibility predictions has been epitope search all four prediction techniques were used to locate antigenic regions in the seven proteins.

Each prediction technique was run with the optimized window sizes. Since there are no general rules to locate the antigenic regions from plots, areas having some sort of peak in the epitope region were considered to match. The VTR, KS, and R parameters predicted correctly 20 of the 31 epitopes which means 65% success ratio. The BP method was much poorer giving only 13 correct regions, 42% success. These figures might be reasonably good for this sort of simple method were there not also a high number of false positives. In Figure 1 we have included also the antigenic index,[19] which is specially made for epitope search. However, it was most often indicating some 60% of the sequence as highly antigenic, thus we did not consider that method at all

Hydropathy profiles have generally been used for searching epitopes. The method of Hopp and Woods,[33] perhaps the most often used prediction technique for this purpose, was used to analyze the same proteins. There were 21 correctly predicted sites indicating no difference in accuracy to flexibility techniques. Because of the vague nature of the flexibility we could not calculate the ratios of correctly and wrongly predicted regions. Anyhow, it could be noted that by far the best methods for searching epitopes among the highest peaks in predictions are VTR and KS. They also predicted fewer false epitopes. The new parameters were better because they separated the peaks more clearly, which makes the interpretation of the results clearer and more accurate. The hydropathy predictions were made with program HYDRO.[34] Note that the hydrophilic regions are pointing down in the Hopp and Woods[33] prediction.

Flexibility Indices

The functional properties of a molecule are a compromise between flexibility and rigidity. The correlation between averaged flexibilities and protein thermal stability has been verified with flexibility indices calculated from the normalized *B*-values of KS.[2] Here we used VTR parameters to calculate also *F* indices. The values determined with KS parameters are shown for comparison. The differences in KS results to those previously published are due to a minor error in the routine for calculating F_7 in the previous work. Several groups of enzymes studied (Table III) indicated that the correlation to protein stability was even clearer with the new parameters. Indices calculated with VTR parameters show correlation also in alanine dehydrogenases, glucoamylases, serine proteases, and phosphoglycerate kinases, but not with those of KS. The flexibility indices are comparable for proteins having similar function and folding. Because they do not take into

TABLE III. Flexibility Indices of Some Proteins*

Source	Temperature: Optimum	Temperature: Stability	Our parameters: T_m	Our parameters: F	Our parameters: F_9	KS parameters: F	KS parameters: F_9	Reference
Alanine dehydrogenase								
Bacillus sphaericus IFO3525		50%, 65°C, 5 min		1.0041	1.0035	0.9897	0.9897	35
Bacillus stearothermophilus IFO12550		50%, 65°C, 5 min		0.9958	0.9964	0.9903	0.9905	35
α-Amylase								
Bacillus subtilis 168		30%, 65°C, 10 min		1.0551	1.0548	1.0081	1.0080	
Streptomyces griseus IMRU 3570	42			1.0223	1.0217	1.0002	0.9998	36, 37
Bacillus amyloliquefaciens	50–60			1.0470	1.0484	1.0051	1.0054	
Aspergillus oryzae		70%, 50°C, 30 min		1.0125	1.0109	0.9978	0.9972	
B. stearothermophilus ATCC 12980	80			1.0178	1.0187	0.9957	0.9963	
B. stearothermophilus NZ-3		50%, 90°C, 2 h		1.0186	1.0196	0.9967	0.9973	
Bacillus licheniformis NCIB 8061		100%, 90°C, 2h		1.0295	1.0304	0.9971	0.9976	
β-Amylase								
Bacillus circulans NCIB 11033	50	77%, 57°C, 1 h		1.0400	1.0411	1.0044	1.0050	38
Clostridium thermosulfurogenes ATCC 33743		100%, 70°C, 1 h		1.0048	1.0055	0.9936	0.9939	39
β-Glucanase								
B. amyloliquefaciens		50%, 70°C, 4 min		1.0227	1.0190	0.9969	0.9954	40
Bacillus macerans		50%, 70°C, 9 min		1.0106	1.0124	0.9956	0.9950	40
Cyclodextrin glycosyltransferase								
Klebsiella pneumoniae M5a1		100%, 45°C, 15 min (+Ca)		1.0405	1.0395	1.0065	1.0058	
B. macerans IAM 1243	60	90%, 50°C, 15 min		1.0226	1.0221	1.0026	1.0026	
Bacillus circulans ATCC 21783	45	100%, 65°C, 30 min (+Ca)		1.0202	1.0198	0.9979	0.9978	41, 42
Ferredoxin								
Clostridium acidi-urici		22%, 70°C, 2h		0.9970	1.0048	0.9651	0.9702	
Clostridium tartarivorum		53%, 70°C, 2h		0.9622	0.9659	0.9659	0.9660	
Clostridium thermosaccharolyticum		90%, 70°C, 2 h		0.9625	0.9663	0.9677	0.9681	
Glucoamylase								
Schizosaccharomycopsis fibuligera HUT 7212	50			1.0475	1.0482	1.0068	1.0071	43
Schwanniomyces occidentalis ATCC 26076	52			1.0275	1.0279	0.9990	0.9991	44
Aspergillus awamori	70			1.0173	1.0184	1.0067	1.0068	45, 46
Inorganic pyrophosphatase								
Saccharomyces cerevisiae		20%, 50°C, 5 min		1.0405	1.0420	1.0044	1.0038	
Escherichia coli		35%, 90°C, 5 min		1.0273	1.0245	0.9921	0.9912	
Lactate dehydrogenase								
Lactobacillus casei DMS 20011		100%, 60°C, 5 min		1.0153	1.0143	0.9923	0.9917	
Bacillus psychrosaccharolyticus DSM 6	40/35			1.0144	1.0107	0.9904	0.9890	47, 48
Bacillus megaterium		100%, 48°C, 30 min		1.0130	1.0117	0.9906	0.9898	48, 49
B. subtilis X1	50–60	100%, 55°C, 30 min		1.0099	1.0108	0.9897	0.9898	
Bacillus caldotenax YT-G	60/70	100%, 65°C, 30 min		1.0110	1.0090	0.9870	0.9865	
Bacillus caldolyticus		100%, 70°C, 30 min		1.0037	1.0020	0.9805	0.9794	50, 51
B. stearothermophilus NCIB 8924	55/60–70	100%, 75°C, 30 min		1.0021	1.0003	0.9796	0.9786	
Thermus caldophilus GK 24		100%, 90°C, 60 min		0.9992	1.0015	0.9816	0.9821	52
Neutral protease								
B. amyloliquefaciens				1.0407	1.0449	1.0128	1.0134	53
B. subtilis				1.0384	1.0418	1.0143	1.0154	54
B. cereus DSM 3101		75%, 65°C, 20 min		1.0292	1.0295	1.0045	1.0042	55, 56
B. stearothermophilus CU21		80%, 65°C, 30 min		1.0232	1.0252	0.9957	0.9965	57, 58
B. caldolyticus YT-p	77	70%, 70°C, 30 min		1.0215	1.0234	0.9958	0.9961	59
Bacillus thermoproteolyticus		30%, 90°C, 30 min		1.0260	1.0273	0.9995	0.9991	58, 60
B. stearothermophilus MK232		45%, 90°C, 30 min		1.0260	1.0272	0.9998	0.9994	61, 62
Phosphoglycerate kinase								
S. cerevisiae			55	1.0838	1.0316	1.0004	0.9997	4, 63
Thermus thermophilus HB-8			>90	1.0814	1.0330	0.9898	0.9898	4, 64
Serine protease								
B. amyloliquefaciens	50–75			1.0240	1.0353	1.0041	1.0049	65, 66
Thermoactinomyces vulgaris	60–85	50%, 55°C, 40 min		1.0271	1.0271	1.0074	1.0076	67, 68
Thermus aquaticus YT-1	80	85%, 80°C, 3 h		1.0202	1.0197	1.0050	1.0041	69, 70

*If reference is not given it is mentioned in Vihinen.[2]

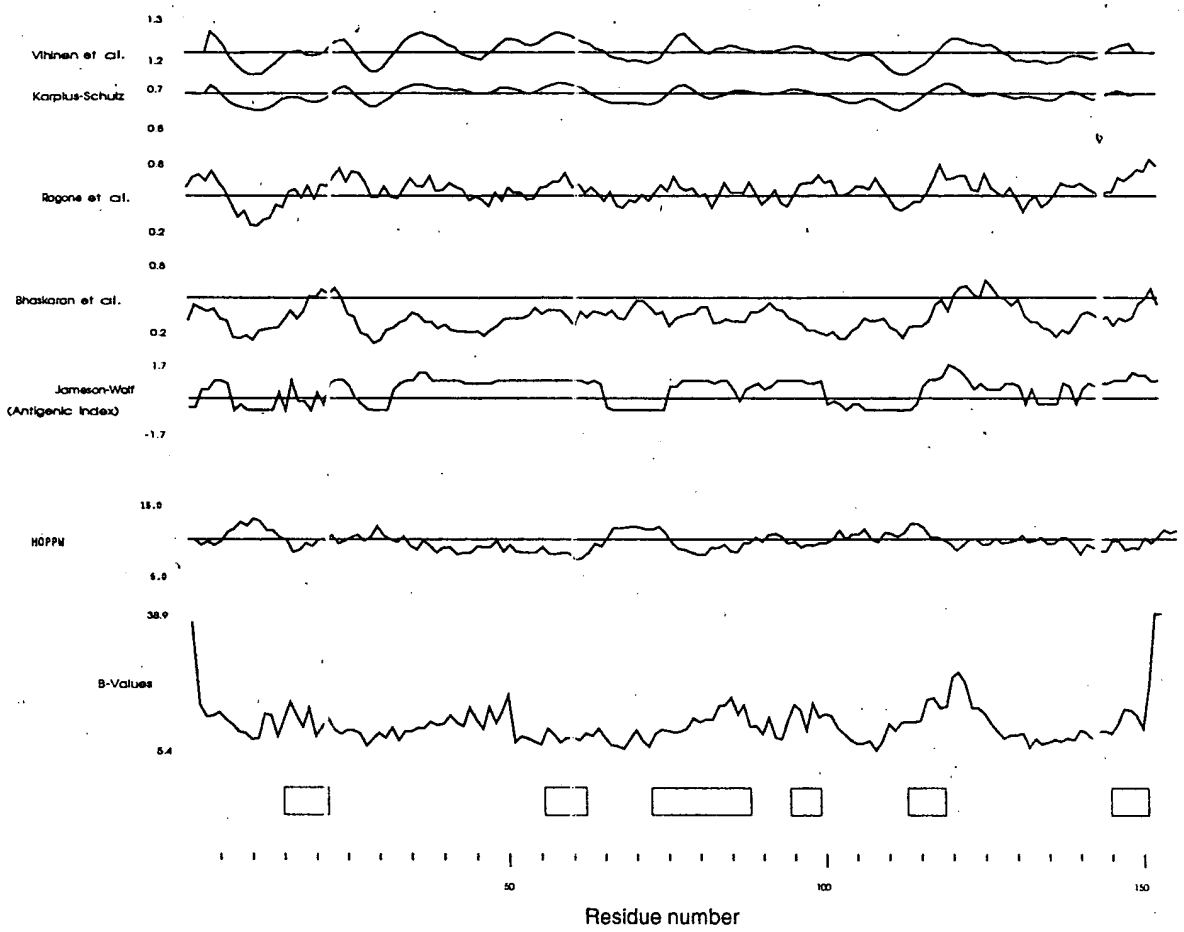

Fig. 1. Flexibility predictions, antigenic index, and experimental *B*-values of backbone atoms in sperm whale myoglogin (1mbo). The flexibility predictions were obtained with optimized prediction windows. Continuous epitopes are indicated with open boxes.

account all the stabilizing forces some discrepancy has been noticed.[2] This can still be seen in the case of the most stable α-amylases, ferredoxins, and neutral proteases. Ferredoxins with extra stabilizing ionic bonds have been discussed previously.[2] In neutral proteases the higher stability is presumably gained by the extra Ca^{2+} binding site. F_9 indices were determined also with KS parameters, but the results are not shown here because the difference to F_7 values were not higher than 0.0005, usually much less.

The flexibility indices calculated with the VTR parameters had more pronounced correlation to stability data than the KS parameters. Our values show correlation even when those of KS fail. This is presumably due to two reasons. Our structural database is larger. We also used the backbone information instead of C_α atoms. All the atoms in residues were not used because flexibility of side chains does not mean that the backbone is also flexible and the flexibility of the protein backbone is typical for surface regions and epitopes. Side chains can be rather mobile although the backbone is rigid. Because one of the major applications of the method will be to search for epitopes and mobile exposed regions only the backbone data were used. Another reason was that often the data for side chains are missing or are poorly determined.

CONCLUSIONS

New parameters were determined for prediction of protein flexibility. The applicability was studied by comparing atomic temperature factors of crystallographically determined proteins to predictions. The VTR and KS parameters were clearly the best. We would suggest the use of a prediction window of 9 residues and VTR parameters because they gave slightly better correlation on a test set of 38 proteins, because they separate flexible regions more clearly on plots, and because they gave much better correlation when used in flexibility indices. It seems that the accuracy of the sliding window technique is approaching its limit and it might be difficult to improve it significantly. The same sort of limits have

also been met in secondary structural predictions where the average accuracy has been for a decade about the same in spite of numerous new methods.[71,72] These limits in prediction techniques are presumably due to intrinsic limitations of the statistical methods, which cannot take into account all the different features of complicated protein structures. Somewhat improved predictions might be obtained with neural nets and other knowledge based systems.

Flexibility profiles can be useful in several ways. When joined with sequence analysis and structural predictions, they can add to our understanding of proteins. In addition to being used for epitope searches, flexibility calculations can be applied in studies concerning sequence and structural similarity, molecular modeling, and protein engineering.

ACKNOWLEDGMENTS

Drs. Lennart Nilsson and Adrian Goldman are thanked for critical reading of the manuscript.

REFERENCES

1. Päivärainen, A. I. "Solvent-Dependent Flexibility of Proteins and Principles of Their Function." Reidel, Dordrecht, 1985.
2. Vihinen, M. Relationship of protein flexibility to thermostability. Prot. Eng. 1:477–480, 1987.
3. Fontana, A. Structure and stability of thermophilic enzymes. Studies on thermolysin. Biophys. Chem. 29:181–193, 1988.
4. Varley, P. G. Pain, R. H. Relation between stability, dynamics and enzyme activity in 3-phosphoglycerate kinases from yeast and *Thermus thermophilus*. J. Mol. Biol. 220: 531–538, 1991.
5. Artymiuk, P. J., Blake, C. C. F., Grace, D. E. P., Oatley, S. J. Phillips, D. C., Sternberg, M. J. E. Crystallographic studies of the dynamic properties of lysozyme. Nature (London) 280:563–568, 1979.
6. Bennett, W. S., Steitz, T. A. Glucose-induced conformational change in yeast hexokinase. Proc. Natl. Acad. Sci. U.S.A. 75:4848–4852, 1978.
7. Farnum, M. F., Magde, D., Howell, E. E., Hirai, J. T., Warren, M. S., Grimsley, J. K., Kraut, J. Analysis of hydride transfer and cofactor fluorescence decay in mutants of dihydrofolate reductase: possible evidence for participation of enzyme molecular motions in catalysis. Biochemistry 30:11567–11579, 1991.
8. Huston, E. E., Grammer, J. C., Yount, R. G. Flexibility of myosin heavy chain: direct evidence that the region containing SH1 and SH2 can move 10 Å under influence of nucleotide binding. Biochemistry 27:8945–8952, 1988.
9. Novotny, J., Handschumacher, H., Haber, E., Bruccoleri, R. E., Carlson, W. B., Fanning, D. W., Smith, J. A., Rose, G. D. Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). Proc. Natl. Acad. Sci. U.S.A. 83:226–230, 1986.
10. Tanaka, T., Kato, H., Nishioka, T., Oda, J. Mutational and proteolytic studies of a flexible loop in glutathione synthetase from *Escherichia coli* B: The loop and arginine 233 are critical for the catalytic reaction. Biochemistry 31: 2259–2265, 1992.
11. Barford, D., Johnson, L. N. The allosteric transition of glycogen phosphorylase. Nature (London) 340:609–616, 1989.
12. Karplus, P. A., Schulz, G. E. Prediction of chain flexibility in proteins, Naturwissenschaften 72:212–213, 1985.
13. Bhaskaran, R., Ponnuswamy, P.K. Positional flexibilities of amino acid residues in globular proteins. Int. J. Peptide Prot. Res. 32:241–255, 1988.
14. Ragone, R., Facchiano, F., Facchiano, A., Facchiano, A. M. Colonna, G. Flexibility plot of proteins. Prot. Eng. 2:497–504, 1989.
15. Enomaa, N., Heiskanen, T., Halila, R., Sormunen, R., Seppälä, R., Vihinen, M., Peltonen, L. Human aspartylglucosaminidase. A biochemical and immuno-cytochemical characterization of the enzyme in normal and aspartylglucosaminuria fibroblasts. Biochem. J. 286:613–618, 1992.
16. Parker, J. M. R., Guo, D., Hodges, R. S. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry 25:5425–5432, 1986.
17. Hopp, T. P. Use of hydrophilicity plotting procedures to identify protein antigenic segments and other interaction sites. Methods Enzymol. 178:571–585, 1989.
18. Krchnak, V., Mach, O., Maly, A. Computer prediction of potential immunogenic determinants from protein amino acid sequence. Anal. Biochem. 165:200–207, 1987.
19. Jameson, B. A., Wolf, H. The antigenic index: A novel algorithm for predicting antigenic determinants. Comput. Appl. Biosci. 4:181–186, 1988.
20. Stern, P. S. Predicting antigenic sites in proteins. Trends Biotech. 9:163–169, 1991.
21. Boberg, J., Salakoski, T., Vihinen, M. Selection of a representative set of structures from Brookhaven Protein Data Bank. Proteins 14:265–276, 1992.
22. Menendez-Arias, L., Argos, P. Engineering protein thermal stability. Sequence statistics point to residue substitutions in α-helices. J. Mol. Biol. 206:397–406, 1989.
23. Devereux, J., Haeberli, P., Smithies, O. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12:387–395, 1984.
24. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:537–542, 1977.
25. Claverie, J.-M., Daulmerie, C. Smoothing profiles with sliding windows: better to wear a hat! CABIOS 7:113–115, 1991.
26. von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J. Mol. Biol. 225:487–494, 1992.
27. Vihinen, M., Eurasto, A., Luostarinen, P., Nevalainen, O. MULTICOMP, a program package for multiple sequence comparison. CABIOS 8:35–38, 1992.
28. Vihinen, M., Simultaneous comparison of several sequences. Methods Enzymol. 183:447–456, 1990.
29. Westhof, E., Altschuh, D., Moras, D., Bloomer, A. C., Mondragon, A., Klug, A., Van Regenmortel, M. H. V. Correlation between segmental mobility and the location of antigenic determinants in proteins. Nature (London) 311:123–126, 1984.
30. Atassi, M. Z. Antigenic structures of proteins. Their determination has revealed important aspects of immune recognition and generated strategies for synthetic mimicking of protein binding sites. Eur. J. Biochem. 145:1–20, 1984.
31. Dyall-Smith, M. L., Lazdins, I., Tregear, G. W., Holmes, I. H. Location of major antigenic sites involved in rotavirus serotype-specific neutralization. Proc. Natl. Acad. Sci. U.S.A. 83:3465–3468, 1986.
32. Argos, P., Fuller, S. D. A model for the hepatitis B virus core protein: prediction of antigenic sites and relationship to RNA virus capsid proteins. EMBO J. 7:819–824, 1988.
33. Hopp, T. P., Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. 78:3824–3828, 1981.
34. Vihinen, M. and Torkkila, E., HYDRO, a program for protein hydropathy prediction. Comput. Meth. Progr. Biomed. 41:121–129, 1993.
35. Kuroda, S., Tanizawa, K., Sakamoto, Y., Tanaka, H., Soda, K. Alanine dehydrogenases from two *Bacillus* species with distinct thermostabilities: molecular cloning, DNA and protein sequence determination, and structural comparison with other NAD(P)$^+$-dependent dehydrogenases. Biochemistry 29:1009–1015, 1990.
36. Garcia-Conzalez, M. D., Martin, J. F., Vigal, T., Liras, P. Characterization, expression in *Streptomyces lividans*, and processing of the amylase of *Streptomyces griseus* IMRU 3570: two different amylases are derived from the same gene by an intracellular processing mechanism. J. Bacteriol. 173:2451–2458, 1991.

37. Vigal, T., Gil, J. A., Daza, A., Garcia-Gonzalez, M. D., Martin, J. F. Cloning, characterization and expression of an α-amylase gene from *Streptomyces griseus* IMRU3570. Mol. Gen. Genet. 225:278–288, 1991.
38. Siggens, K. W. Molecular cloning and characterization of the beta-amylase gene from *Bacillus circulans*. Mol. Microbiol. 1:86–91, 1987.
39. Kitamoto, N., Yamagata, H., Kato, T., Tsukagoshi, N., Udaka, S. Cloning and sequencing of the gene encoding thermophilic β-amylase of *Clostridium thermosulfurogenes*. J. Bacteriol. 170:5848–5854, 1988.
40. Borriss, R., Olsen, O., Thomsen, K. K., von Wettstein, D. Hybrid *Bacillus* endo-(1-2,1-4)-β-glucanases: construction of recombinant genes and molecular properties of the gene products. Carlsberg Res. Commun. 54:41–54, 1989.
41. Nakamura, N., Horikoshi, K. Purification and properties of cyclodextrin glycosyltransferase of an alkalophilic *Bacillus* sp., Agric. Biol. Chem. 40:935–941, 1976.
42. Kaneko, T., Hamamoto, T., Horikoshi, K. Molecular cloning and nucleotide sequence of the cyclomaltodextrin glucanotransferase gene from the alkalophilic *Bacillus* sp. strain no. 38-2. J. Gen. Microbiol. 134:97–105, 1988.
43. Yamashita, I., Itoh, T., Fukui, S. Cloning and expression of the *Saccharomycopsis fibuligera* glucoamylase gene in *Saccharomyces cerevisiae*. Appl. Microbiol. Biotechnol. 23: 130–133, 1985.
44. Dohmen, R. J., Strasser, A. W. M., Dahlems, U. M., Hollenberg, C. P. Cloning of *Schwanniomyces occidentalis* glucoamylase gene (GAM1) and its expression in *Saccharomyces cerevisiae*. Gene 95:111–115, 1990.
45. Nunberg, J. H., Meade, J. H., Cole, G., Lawyer, F. C., McCabe, P., Schweickart, V., Tal, R., Wittman, V. P., Flatgaard, J. E., Innis, M. A. Molecular cloning and characterization of the glucoamylase gene of *Aspergillus awamori*. Mol. Cell. Biol. 4:2306–2315, 1984.
46. Evans, R., Ford, C., Sierks, M., Nikolov, Z., Svensson, B. Activity and thermal stability of genetically truncated forms of *Aspergillus* glucoamylase. Gene 91:131–134, 1990.
47. Schlatter, D., Kriech, O., Suter, F., Zuber, H. The primary structure of the psychrophilic lactate dehydrogenase from *Bacillus psychrosaccharolyticus*. Biol. Chem. Hoppe-Seyler 368:1435–1446, 1987.
48. Zülli, F., Schneiter, R., Urfer, R., Zuber, H. Engineering thermostability and activity of lactate dehydrogenases from *Bacilli*. Biol. Chem. Hoppe-Seyler 372:363–372, 1991.
49. Stangl, D., Widerkehr, F., Suter, F., Zuber, H. The complete amino-acid sequence of the mesophilic L-lactate dehydrogenase from *Bacillus megaterium*. Biol. Chem. Hoppe-Seyler 368:1157–1166, 1987.
50. Zülli, F., Weber, H., Zuber, H. Nucleotide sequences of lactate dehydrogenase genes from the thermophilic bacteria *Bacillus stearothermophilus, B. caldolyticus* and *B. caldotenax*. Biol. Chem. Hoppe-Seyler 368:1167–1177, 1987.
51. Zülli, F., Weber, H., Zuber, H. Analysis of structural elements responsible for the differences in thermostability and activation by fructose 1,6-bisphosphate in lactate dehydrogenases from *B. stearothermophilus* and *B. caldolyticus* by protein engineering. Biol Chem. Hoppe-Seyler 371: 655–662, 1990.
52. Kunai, K., Machida, M., Matsuzawa, H., Ohta, T. Nucleotide sequence and characteristics of the gene for L-lactate dehydrogenase of *Thermus caldophilus* GK24 and the deduced amino-acid sequence of the enzyme. Eur. J. Biochem. 160:433–440, 1986.
53. Vasantha, N., Thompson, L. D., Rhodes, C., Banner, C., Nagle, J., Filpula, D. Genes for alkaline protease and neutral protease from *Bacillus amyloliquefaciens* contain a large open reading frame between the regions coding for signal sequence and mature protein. J. Bacteriol. 159:811–819, 1984.
54. Yang, M. Y., Ferrari, E., Henner, D. J. Cloning of the neutral protease gene of *Bacillus subtilis* and the use of the cloned gene to create in vitro-derived deletion mutants. J. Bacteriol. 160:15–21, 1984.
55. Sidler, W., Kumpf, B., Peterhans, B., Zuber, H. A neutral proteinase produced by *Bacillus cereus* with high sequence homology to thermolysin: Production, isolation and characterization. Appl. Microbiol. Biotechnol. 25:18–23, 1986.
56. Sidler, R., Niederer, E., Suter, F., Zuber, H. The primary structure of *Bacillus cereus* neutral proteinase and comparison with thermolysin and *Bacillus subtilis* neutral proteinase. Biol. Chem. Hoppe-Seyler 367:643–657, 1986.
57. Takagi, M., Imanaka, T., Aiba, S. Nucleotide sequence and promoter region for the neutral protease gene from *Bacillus stearothermophilus*. J. Bacteriol. 163:824–831 1985.
58. Imanaka, T., Shibazaki, M., Takagi, M. A new way of enhancing the thermostability of proteases. Nature (London) 324:695–697, 1986.
59. van den Burg, B., Enequist, H. G., van der Haar, M. E., Eijsink, V. G. H., Stulp, B. K., Venema, G. A highly thermostable neutral protease from *Bacillus caldolyticus*: Cloning and expression of the gene in *Bacillus subtilis* and characterization of the gene product. J. Bacteriol. 173: 4107–4115, 1991.
60. Titani, K., Hermodson, M. A., Ericsson, L. H., Walsh, K. A., Neurath, H. Amino-acid sequence of thermolysin. Nature (London) 238:35–37, 1972.
61. Kubo, M., Imanaka, T. Cloning and nucleotide sequence of the highly thermostable neutral protease from *Bacillus stearothermophilus*. J. Gen. Microbiol. 134:1883–1892, 1988.
62. Kubo, M., Murayama, K., Seto, K., Imanaka, T. Highly thermostable neutral protease from *Bacillus stearothermophilus*. J. Ferment. Technol. 66:13–17, 1988.
63. Hilzeman, R. A., Hagie, F. E., Hayflick, J. A., Chen, C. Y., Seeburg, P. H., Derynck, R. The primary structure of the *Saccharomyces cerevisiae* gene for phosphoglycerate kinase. Nucleic Acids Res. 10:7791–7808, 1982.
64. Bowen, D., Littlechild, J. A., Fothergill, J. E., Watson, H. C., Hall, L. Nucleotide sequence of the phosphoglycerate kinase gene from the extreme thermophile *Thermus thermophilus*. Biochem. J. 254:509–517, 1988.
65. Wells, J. A., Ferrari, E., Henner, D. J., Estell, D. A., Chen, E. Y. Cloning, sequencing and secretion of *Bacillus amyloliquefaciens* subtilisin in *Bacillus subtilis* Nucleic Acids Res. 11:7911–7925, 1983.
66. Wells, J. A., Powers, D. B. *In vivo* formation and stability of engineered disulfide bonds in subtilisin. J. Biol. Chem. 261:6564–6570, 1986.
67. Kleine, R. Properties of thermitase, a thermostable serine protease from *Thermoactinomyces vulgaris*. A. Biol. Med. Chem. Germ. 41:89–102, 1982.
68. Meloun, B., Baudys, M., Kostka, V., Hansdorf, G., Frömmel, C., Höhne, W. E. Complete primary structure of thermitase from *Thermoactinomyces vulgaris* and its structural features related to the subtilisin-type proteinases. FEBS Lett. 183:195–199, 1985.
69. Kwon, S.-T., Terada, I., Matsuzawa, T., Ohta, Y. Nucleotide sequence of the gene for aqualysin I (a thermophilic alkaline serine protease) of *Thermus aquaticus* YT-1 and characteristics of the deduced primary structure of the enzyme. Eur. J. Biochem. 173:491–497, 1988.
70. Matsuzawa, H., Tokugawa, K., Hamaoki, M., Mizoguchi, M., Taguchi, H., Terada, I., Kwon, S.-T., Ohta, T. Purification and characterization of aqualysin I (a thermophilic alkaline serine protease) produced by *Thermus aquaticus* YT-1. Eur. J. Biochem. 171:441–447, 1988.
71. Kabsch, W., Sander, C. How good are predictions of protein secondary structure? FEBS Lett. 155:179–182, 1983.
72. Rost, B., Schneider, R., Sander, C. Progress in protein structure prediction? Trends Biochem. Sci. 18 120–123, 1993.

J. Mol. Biol. (1992) 225, 1049–1063

Hybrid System for Protein Secondary Structure Prediction

Xiru Zhang, Jill P. Mesirov and David L. Waltz

Thinking Machines Corporation
245 First Street
Cambridge, MA 02142, U.S.A.

(Received 11 October 1991; accepted 10 February 1992)

We have developed a hybrid system to predict the secondary structures (α-helix, β-sheet and coil) of proteins and achieved 66·4% accuracy, with correlation coefficients of C_{coil} = 0·429, C_{α} = 0·470 and C_{β} = 0·387. This system contains three subsystems ("experts"): a neural network module, a statistical module and a memory-based reasoning module. First, the three experts independently learn the mapping between amino acid sequences and secondary structures from the known protein structures, then a Combiner learns to combine automatically the outputs of the experts to make final predictions. The hybrid system was tested with 107 protein structures through k-way cross-validation. Its performance was better than each expert and all previously reported methods with greater than 0·99 statistical significance. It was observed that for 20% of the residues, all three experts produced the same but wrong predictions. This may suggest an upper bound on the accuracy of secondary structure predictions based on local information from the currently available protein structures, and indicate places where non-local interactions may play a dominant role in conformation. For 64% of the residues, at least two experts were the same and correct, which shows that the Combiner performed better than majority vote. For 77% of the residues, at least one expert was correct, thus there may still be room for improvement in this hybrid approach. Rigorous evaluation procedures were used in testing the hybrid system, and statistical significance measures were developed in analyzing the differences among different methods. When measured in terms of the number of secondary structures (rather than the number of residues) that were predicted correctly, the prediction produced by the hybrid system was also better than those of individual experts.

Keywords: protein secondary structure prediction; hybrid system; neural networks; memory-based reasoning; statistical methods

1. Introduction

Determining the mapping between amino acid sequences and secondary structures (α helix, β sheet, etc.) is an important step towards our understanding of how protein sequences specify their overall structures and functions. Currently the main technique to determine protein structures is X-ray crystallography, which is a slow and often difficult process. On the other hand, the database of known protein sequences is growing very rapidly. Thus, it is increasingly important to develop computational approaches to determine automatically (predict) the structures of proteins whose sequences are known. The correct prediction of secondary structures can contribute significantly towards this goal. For example, the knowledge of secondary structures can provide a good starting point and reduce the search space in simulation of protein folding by molecular dynamics (Levitt, 1983) or lattice models (Skolnick & Kolinski, 1990), or can be used in predicting higher order structures (e.g. super secondary structures (Taylor & Thornton, 1984), domains (Rathrop *et al.*, 1987)).

Many algorithms have been developed for protein secondary structure prediction. One of the first efforts was made by Chou & Fasman (1974). Different implementations of their algorithm have all attained about a 50 to 60% level of accuracy in predicting the location of α helices, β strands and "coil" (i.e. anything other than helix or strand) in a protein sequence. Garnier, Osguthorpe & Robson's algorithm (Garnier *et al.*, 1978) is about 58% accurate for this task. More recently, their improved algorithm (Gibrat *et al.*, 1987) is 63% accurate. Qian & Sejnowski (1988) used an artificial neural network algorithm to increase the prediction accuracy to 64%. Similar results have also been achieved by other researchers (e.g. Kneller *et al.*, 1990; Holley & Karplus, 1989). Thus there has been about a 6% improvement of prediction accuracy in

0022-2836/92/121049-15 $03.00/0

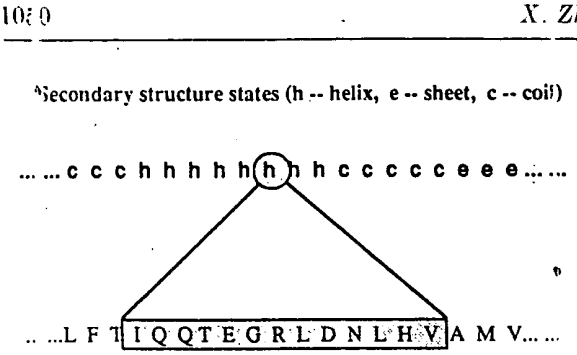

Figure 1. A window is moved along an amino acid sequence to extract correlations between the residues and the secondary structure state of the center residue.

the last 15 to 20 years, which is due to both the improved computational methods and the increase of the known protein structure data. Almost all these algorithms have adopted a "local strategy": moving a "window" (typically covering 7 to 19 residues) along an amino acid sequence and predict the secondary structure state of the center residue in the window according to all the residues inside the window (see Fig. 1). To assess the accuracy of a prediction algorithm for proteins whose structures are not known, it is a common practice to divide the known protein structure database into two separate sets: the "training data set" is used to set the parameters of the algorithm, and the "test data set" is used to test its prediction accuracy. The predictions produced by the existing algorithms, though imperfect, can often show the likelihood or tendency of certain peptide chains to form particular secondary structures. It is also important to know the extent to which the protein structures are determined by "local interactions": interactions among residues adjacent along the polypeptide chain.

Though existing prediction algorithms are all about 60 to 64% accurate for three-state (α-helix, β-sheet, and coil) prediction, they can make incorrect predictions at different places of an amino acid sequence. From the point of view of machine learning (artificial intelligence), secondary structure prediction is an instance of *inductive learning*, generalizing from known examples to solve new problems. Different algorithms may work according to different principles and can generalize in different ways. Therefore, a combination of different algorithms can potentially produce a better prediction than individual ones. Based on this analysis, we developed a hybrid system to predict the secondary structures, which indeed improved the prediction accuracy significantly. Our hybrid system has three different modules ("experts"): a neural network module, a statistical module and a memory-based reasoning module, and a Combiner. The experts were chosen in such a way that they have different mathematical properties. In the training phase, the experts independently learn the mapping between amino acid sequences and secondary structures from the known protein structures; the Combiner learns to combine automatically the outputs of the experts. In the prediction phase, the three experts make predictions separately, then the Combiner takes the predictions from the three experts and makes final predictions. K-way cross-validation was used in evaluating the hybrid system and statistical significance measures were used in comparing different prediction algorithms.

Our experiments showed that (1) the hybrid system had an overall prediction accuracy of 66·4%, which was higher than individual experts and all previously reported algorithms at greater than 0·99 confidence level; (2) the three experts not only had very close overall prediction accuracy, their detailed predictions also agreed with one another much more than with the real structure (i.e. their prediction accuracy); (3) the accuracy of prediction algorithms could change as the test data changes, especially when the test data set was small (e.g. containing 15 protein sequences); (4) for 20% of the residues, all three very different experts produced the same but wrong prediction, suggesting that with the currently available protein structure data, 80% may be the upper bound for the secondary structure prediction accuracy using the local strategy; (5) compared to each expert, the hybrid system also produced better result in terms of the number of secondary structures (rather than the number of residues) that were predicted correctly.

2. Methods and Materials

(a) *The architecture and training of a hybrid system*

Figure 2 shows the overall architecture of our hybrid system. The system contains three "experts", a statistical module, a memory-based reasoning module and a neural network module, and a Combiner. The whole system produces secondary structure predictions as follows: given a set of amino acid sequences (i.e. test data), each expert makes its predictions independently, then the Combiner takes the predictions from the 3 experts and combines them to produce final predictions. At the beginning, the hybrid system learns from the training data set about mappings between amino acid sequences and secondary structures. The training of the whole system involves (1) training the 3 experts and (2) training the Combiner. How each expert is trained and how each makes predictions are discussed in the following sections. In order to train the Combiner, half of the training data is used to train the 3 experts separately, and the outputs of these trained experts on the second half of the training data are recorded. These outputs are then used as inputs to train the Combiner. The reason for dividing the training data set into 2 parts is that the behavior of each expert on training data can be very different from its behavior on the proteins whose structures are unknown; their performance on the data that they are not trained on (the second half of the training set) reflects their behaviors on truly unknown protein structures, which is exactly what the Combiner should know about and be trained on. The training of the experts with half of the training data is done purely for the purpose of training the Combiner. After the training of the Combiner is completed, each

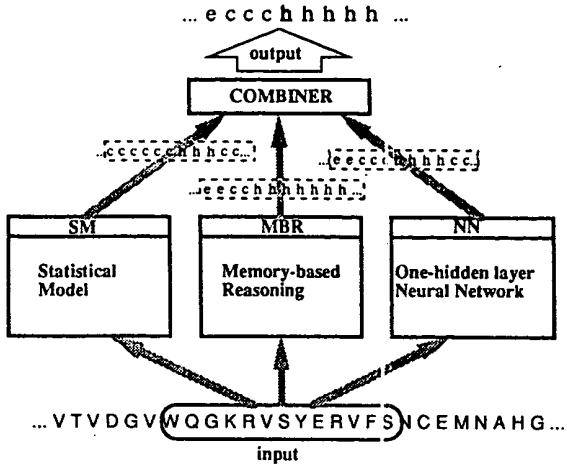

Figure 2. The hybrid system has 3 experts, a statistical module, a memory-based reasoning module and a neural network module. The Combiner combines the outputs of the 3 experts to produce a final output.

expert is trained again with the whole training data set. These trained experts together with the trained Combiner form a trained hybrid system.

(b) *Memory-based reasoning*

Memory-based reasoning (MBR†) (Stanfill & Waltz, 1986) is one expert in our hybrid system. The essential idea of MBR is to use known examples directly in problem solving. For predicting the protein secondary structures, this involves matching each segment (window) of amino acid sequences in the test data set against all the sequences in the training set, finding its "nearest neighbors", and choosing the secondary structure state of the majority of its neighbors as the prediction. Similar approaches have been referred to as the "nearest neighbor method", "exemplar-based reasoning", etc. Levin *et al.* (1986) and Nishikawa & Ooi (1986) called this approach the "homologous method". The key component in this approach is the distance function or metric used to compute the neighbors. The choice of a metric is especially difficult for elements such as amino acids, because there is no linear ordering among the elements, which are often referred to as having "nominal values". Stanfill & Waltz (1986) proposed several distance functions for nominal values in their work on memory-based reasoning. We improved their functions and applied them to protein secondary structures in this work.

Based on the idea of MBR, one distance matrix is computed for each position of the window using the training data set. At window position i, the distance matrix D_i contains the distance between every pair of amino acids at that position. The distance between 2 segments of amino acid sequences $A = a_1 a_2 \ldots a_n$ and $B = b_1 b_2 \ldots b_n$ is defined as:

$$D(A, B) = \sum_{i=1}^{n} D_i(a_i, b_i),$$

where n is the window size, $D_i(a_i, b_i)$ is the distance between amino acids a_i and b_i at position i. The smaller this distance is, the more similar a_i and b_i are in terms of forming secondary structures, and the less effect it has on secondary structures if one is replaced by the other. The distance matrices D_i can be computed from the training data. Assuming there are m secondary structure states s_1, s_2, $\ldots s_m$ and q different amino acids, x^1, x^2, $\ldots x^q$ ($a_i\, b_i \in \{x^1, \ldots x^q\}$), $D_i(a_1, b_i)$ is computed as:

$$D_i(a_i, b_i) = \frac{1}{m} \sum_{j=1}^{m} |p(s_j|a_i) - p(s_j|b_i)| + \frac{1}{m \cdot n \cdot q} \sum_{j=1}^{m} \sum_{k=1}^{n} \sum_{h=1}^{q} |p(s_j|a_i, x_k^h) - p(s_j|b_i, x_k^h)| \quad (1)$$

where x_k^h denotes amino acid x^h at window position k; $p(s_j|a_i)$ is the conditional probability of secondary structure state s_j given that a_i has occurred; it represents the influence on secondary structure s_j by the singleton amino acid at position i. $p(s_j|a_i, x_k^h)$ is the conditional probability of s_j given both a_i and x_k^h have occurred; it represents the influence on s_j by a_i together with its neighbor amino acids. Thus when $p(s_j|a_i) \approx p(s_j|b_i)$ and $p(s_j|a_i, x_k^h) \approx p(s_j|b_i, x_k^h)$, a_i and b_i are similar in determining secondary structures, and $D_i(a_i, b_i)$ should be small, which is exactly what equation (1) yields.

(c) *A statistical method*

A statistical module (SM) is the second expert in our hybrid system. It works as follows: for each secondary structure state s_j, if the conditional probability of s_j given a window of n residues $a_1 \ldots a_n$, $p(s_j|a_1 \ldots a_n)$, is known, then the s_j that has the highest value for this conditional probability is chosen as the prediction for $a_1 \ldots a_n$:

$$\text{Prediction} = \left\{ s_j | \max_j p(s_j|a_1, a_2, \ldots a_n) \right\},$$

$$s_j \in \{\alpha\text{-helix}, \beta\text{-sheet}, \text{coil}\}.$$

According to Bayes Theorem:

$$p(s_j|a_1 \ldots a_n) = \frac{p(s_j) \cdot p(a_1 \ldots a_n|s_j)}{p(a_1 \ldots a_n)}. \quad (2)$$

where $p(s_j)$ is the probability of s_j and $p(a_1 \ldots a_n|s_j)$ is the probability of $a_1 \ldots a_n$ in secondary structure state s_j; $p(a_1 \ldots a_n)$ is the probability of $a_1 \ldots a_n$ in all states. Since we only want to find the largest $p(s_j|a_1 \ldots a_n)$, $p(a_1 \ldots a_n)$ need not be computed. Currently there is not enough protein structure data available for us to compute the frequencies of $a_1 \ldots a_n$ in each state s_j in order to estimate $p(a_1 \ldots a_n|s_j)$. They have to be estimated by some simpler terms. We extend and apply the Bahadur-Lazarsfeld expansion (Bahadur, 1961) here (which only deals with binary variables in its original form). Assuming that $y_1, y_2, \ldots y_n$ are random variables with nominal values, then

$$p(y_1, \ldots, y_n) = \prod_i^n p(y_i) \times \left\{ 1 + \sum_{i<k} Z_{ik} + \sum_{i<k<h} Z_{ikh} + \ldots \right\}. \quad (3)$$

where Z_{ik} is the second order correlation between y_i and y_k:

$$Z_{ik} = \frac{p(y_i, y_k)}{p(y_i)p(y_k)} - 1,$$

† Abbreviations used: MBR, memory-based reasoning; IP, input pattern; SM, statistical module.

and Z_{ikh} is the third order correlation among y_i, y_k and y_h:

$$Z_{ikh} = \left(\frac{p(y_i, y_k, y_h)}{p(y_i)p(y_k)p(y_h)} - 1\right) - \left(\frac{p(y_i, y_k)}{p(y_i)p(y_k)} - 1\right) - \left(\frac{p(y_i, y_h)}{p(y_i)p(y_h)} - 1\right) - \left(\frac{p(y_k, y_h)}{p(y_k)p(y_h)} - 1\right).$$

and so on.

In practice, for the secondary structure prediction problem, we can only estimate up to the second order correlations with the currently available protein structure data. The reliability of these estimates depends on the sample size used. Thus, we postulate the following equation:

$$p(a_1, \ldots, a_n|s_j) \approx \prod_i p(a_i|s_j) \cdot \left[1 + C_f \cdot \sum_{i<k} f_{ik} \cdot \left(\frac{p(a_i, a_k|s_j)}{p(a_i|s_j)p(a_k|s_j)} - 1\right)\right] \quad (4)$$

where f_{ik} is proportional to the size of the sample in which ratio

$$\frac{p(a_i, a_k|s_j)}{p(a_i|s_j)p(a_k|s_j)}$$

is computed, to represent its reliability:

$$f_{ik} = \sqrt{\frac{p(a_i|s_j)p(a_k|s_j)}{(1-p(a_i|s_j))(1-p(a_k|s_j))}}.$$

Some observations about equation (4): (1) Compared with equation (3), correlations among 3 or more residues are ignored. This is due to the limited sample size. This truncation may have an overall positive or negative effect on the contribution from higher-order correlations in the approximation, thus coefficient C_f is introduced to compensate for this. C_f can be experimentally determined. (2) When there are no higher-order correlations among the residues in a window (i.e. they are all independent), $p(a_1, \ldots, a_n|s_j)$ is reduced to $\prod_i p(a_i|s_j)$, which is correct. (3) Information of all C_n^2 possible pairs of residues in a window of size n is used here, whereas in a commonly used statistical method, the GOR III method, only $n-1$ pairs are used. (4) If the pairwise correlation terms are small and the approximation $\log(1+x) \approx x$ is used, we get the following equation:

$$p(a_i, \ldots, a_n|s_j) \approx \prod_i p(a_i|s_j) \cdot e^{C_f \cdot \sum_{i<k} f_{ik} \cdot \left(\frac{p(a_i, a_k|s_j)}{p(a_i|s_j)p(a_k|s_j)} - 1\right)}. \quad (5)$$

This is exactly the form in Lazarsfeld's original expansion (Lazarsfeld, 1961), which he derived from a completely different path. One advantage of equation (5) is that it guarantees that the probability approximation is non-negative, which equation (4) does not do. Equation (5) is the final form of the statistical expert used in this work.

(d) *Artificial neural network*

Artificial neural networks have been used widely in many applications (McClelland & Rumelhart, 1986), including protein secondary structure prediction (Qian & Sejnowski, 1988; Kneller *et al.*, 1990). An artificial neural network usually consists of a large number of simple processing units connected by weighted links. Each unit computes its output by applying an "activation function" to its inputs. The training algorithm used in this work, the Back-propagation algorithm (Rumelhart *et al.*, 1986), works on a particular kind of artificial neural network, a layered, feed-forward network (see Fig. 3), where the processing units are arranged in layers: there is an input layer, an output layer, and one or more "hidden layers" (layers between the input and output layer). A feed-forward network computes its output in the following fashion: first, the input layer is set according to an input pattern; then one layer at a time, from the input to hidden to output layer, the units compute their outputs by applying an activation function to the weighted sum of the outputs from the units at the lower layer. The weights come from the links between the units. The "sigmoid function" is often used in feedforward networks as the unit's activation function:

$$O_{i,j} = \frac{1}{1+e^{-x}}$$

Where O_{ij} is the output of unit j at layer i, and x is the weighted sum of outputs from units at one layer below:

$$x = \sum_k w_{i-1,k}^{i,j} O_{i-1,k}$$

$w_{h,i-1}^{i,j}$ is the weight of the link from unit k at layer $i-1$ to unit j at layer i. This can also be seen as a projection of the network input to a certain direction specified by the weights. Thus, each hidden unit represents a different projection from the multiple dimensional input space to a new space whose dimensionality is determined by the number of hidden units in the network.

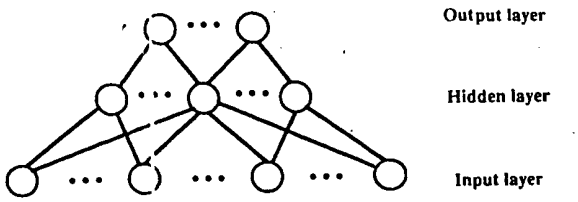

Figure 3. A one-hidden-layer feedforward artificial neural network. The network computes its output based on the values of the units at the input layer.

The Back-propagation algorithm "trains" a layered network by adjusting the link weights of the net using a set of "training examples". Each training example consists of an input pattern and an ideal output pattern that the user wants the network to produce for that input. The weights are adjusted based on the difference between the ideal output and the actual output of the network. This can be seen as a gradient descent process in the weight space. An "epoch training cycle" consists of presenting all training examples once to the network, and then adjusting the weights on the basis of the accumulated errors at the output layer. A number of epoch cycles may be required before the output errors are reduced to an accepted level. After the training is completed, the network can be applied to inputs that are not in the set of training examples. For a new input pattern IP, the trained network tends to produce an output similar to the training example whose input is similar to IP. This can be used for interpolation, approximation, or generalization from examples depending on the goal of the user.

Table 1

Protein structures used in this work

Protein	Code	Subunit	Length	No. H	No. E
Cytochrome c550	155C		134	35	5
Cytochrome B562 (*E. coli*, oxidized)	156B		110	67	0
L-Arabinose-binding protein	1ABP		306	106	20
Actinoxanthin	1ACX		107	0	47
Phospholipase A2	1BP2		123	54	8
Cytochrome c5 (oxidized)	1CC5		83	39	0
Cytochrome c	1CCR		111	44	0
Calcium-binding parvalbumin B	1CPV		108	52	6
Crambin	1CRN		46	20	4
Subtilisin carlsberg (inhibitor)	1CSE		63	11	22
17/112 Ribosomal protein (C-terminal domain)	1CTF		68	35	18
Cytochrome c3	1CY3		118	16	0
Hemoglobin (erythrocruorin, deoxy)	1ECD		136	97	0
Elongation factor tu (domain i)	1ETU		196	78	36
Immunoglobin FAB	1FB4	(H L)	445	11	208
FC fragment (IGG1 class)	1FC1	(A)	206	15	95
Immunoglobin fc and fragment B of protein a complex	1FC2	(C)	43	21	0
Ferrodoxin	1FDX		54	5	4
Flavodoxin	1FX1		147	43	32
Ferredoxin	1FXB		81	10	0
Glucagon (pH 6–pH 7 form)	1GCN		29	14	0
γ Crystallin	1GCR		174	5	77
Glyceraldehyde-3-phosphate dehydrogenase	1GD1	(O)	336	73	95
Glutathione peroxidase	1GP1	(A)	184	43	29
Oxidized high potential iron protein (HIPIP)	1HIP		85	10	9
Hemerythrin (MET)	1HMQ	(A)	113	73	0
Insulin	1INS	(A D)	51	22	3
Leghemoglobin (Acetate, MET)	1LH1		153	107	0
Lysozyme	1LZ1		130	39	10
Myoglobin (DEOXY, pH 8.4)	1MBD		153	113	0
Immunoglobulin FAB fragment (MC/PC603)	1MCP	(H L)	442	8	211
Melittin	1MLT	(A)	26	22	0
Neurotoxin B	1NXB		62	0	26
Pseudoazurin	1PAZ		120	17	44
Plastocyanin	1PCY		99	4	35
Hydroxybenzoate hydroxylase	1PHH		394	119	96
Calcium-free phospholipase A2	1PP2	(L)	133	48	8
Avian pancreatic polypeptide	1PPT		36	18	0
Rhodanese	1RHD		293	81	32
Ribonuclease A	1RN3		124	22	48
Ribonuclease T1 isozyme	1RNT		104	17	28
Subtilisin BPN	1SBT		275	83	49
Trypsin (SGT)	1SGT		240	21	77
Scorpion neurotoxin (variant 3)	1SN3		65	8	12
Trypsinogen complex with porcine pancreatic secretory	1TGS	(I)	57	9	11
Triose phosphate isomerase	1TIM	(A)	248	106	42
Tonin	1TON		238	10	71
Ubiquitin	1UBQ		76	12	24
α-Bungarotoxin	2ABX	(A)	74	0	4
Actinidin (sulfhydryl proteinase)	2ACT		218	56	40
Acid proteinase, penicillopepsin	2APP		323	30	147
Acid proteinase (rhizopuspepsin)	2APR		325	26	146
Azurin (oxidized)	2AZA	(B)	129	13	41
Cytochrome B5 (oxidized)	2B5C		85	21	21
Carbonic anhydrase form B (carbonate dehydratase)	2CAB		256	17	79
Cytochrome c'	2CCY	(A)	127	90	0
Cytochrome c3	2CDV		107	27	10
Chymotrypsinogen A	2CGA	(A)	245	18	79
Chymotrypsin inhibitor 2 (CI-2)	2CI2		65	11	14
Concanavalin A	2CNA		237	4	103
Cytochrome P450CAM (camphor monooxygenase)	2CPP		405	180	41
Citrate synthase	2CTS		437	257	6
Cytochrome c peroxidase	2CYP		293	134	16
Gene 5 DNA binding protein	2GN5		87	0	4
Hemoglobin (deoxy)	2HHB	(A B)	287	197	0
Hemoglobin V (CYANO,MET)	2LHB		149	100	0
Lysozyme	2LZM		164	109	14
Cytoplasmic malate dehydrogenase	2MDH	(A B)	649	213	110

Table 1 *(continued)*

Protein	Code	Subunit	Length	No. H	No. E
CD. ZN Metallothionein (isoform II)	2MT2		61	0	0
Ovomucoid third domain	2OVO		56	10	9
Prealbumin (human plasma)	2PAB	(A)	114	8	59
Proteinase K	2PRK		279	66	60
Staphylococcal nuclease complex	2SNS		141	26	28
CU.ZN Superoxide dismutase	2SOD	(B)	151	0	54
Streptomyces subtilisin inhibitor	2SSI		107	17	26
Satellite tobacco necrosis virus	2STV		184	18	82
Tomato bushy stunt virus	2TBV	(C)	321	4	112
Cytochrome c551 (oxidized)	351C		82	38	0
Adenylate kinase	3ADK		194	106	25
Bacteriochlorophyll	3BCL		356	57	170
Cytochrome c2 (reduced)	3C2C		112	44	0
Native elastase	3EST		251	13	82
Ferredoxin	3FXC		98	7	15
Catabolite gene activator protein–cyclic AMP complex	3GAP	(A)	208	64	21
Glutathione reductase, oxidized form (E)	3GRS		461	132	111
Calcium-binding protein	3ICB		75	43	0
Phosphoglycerate kinase complex with ATP	3PGK		415	143	46
Phosphoglycerate mutase DE-phospho enzyme	3PGM		230	69	15
Rat mast cell protease II	3RP2	(A)	237	12	83
Rubredoxin	3RXN		52	0	8
Wheat germ agglutinin (isolectin 2)	3WGA	(B)	171	16	16
TRP aporepressor	3WRP		101	77	0
APO-liver alcohol dehydrogenase	4ADH		374	79	77
Aspartate carbamoyltransferase	4ATC	(A B)	463	133	65
Carboxypeptidase Aα (COX) complex	4CPA	(I)	37	0	6
Dihydrofolate reductase complex	4DFR	(B)	159	29	56
Ferredoxin	4FD1		106	18	14
Flavodoxin (semiquinone form)	4FXN		138	47	29
Lactate dehydrogenase APO enzyme M4	4LDH		333	111	37
Trypsin inhibitor	4PTI		58	8	14
β Trypsin, diisopropylphosphoryl inhibited	4PTP		234	16	72
Southern bean mosaic virus coat protein	4SBV	(C)	222	32	72
Thermolysin complex	4TLN		316	117	54
Troponin C	4TNC		160	101	6
Carboxypeptidase Aα (COX)	5CPA		307	111	50
Catalase	7CAT	(A)	498	137	71
Papain CYS-25 oxidized	9PAP		212	49	36
Total		113	19,861	5324	4098

If there is more than 1 subunit in a protein, column Subunit indicates which subunit(s) was used. Length indicates the number of residues in the protein sequences used. No. H indicates the number of residues in α-helix; No. E indicates the number of residues in β-sheet. There are 107 proteins in this Table, with 113 subunits, 19,861 residues.

(e) *Database*

A database of 107 proteins was selected from Brookhaven Protein Data Bank. It contains 19,861 residues, 113 subunits. All sequences (subunits) are less than 50% homologous with one another. The DSSP program (Kabsch & Sander, 1983*a*) was used to assign the secondary structure state of each residue. The DSSP program assigned 7 states, B, E, G, H, S, T and "the rest" to the residues in our database. For the purpose of this work, H was considered α helix, E was considered β sheet, and the rest were considered coil. Table 1 lists names of all the proteins in our database.

(f) *Prediction accuracy measurements*

In this work, we adopted the commonly used definition of prediction accuracy, which is the percentage of correctly predicted residues for the 3 types of secondary structures:

$$Q_3 = \frac{q_\alpha + q_\beta + q_{\text{coil}}}{N},$$

where N is the total number of residues in the test data sets, q_s is the number of residues of secondary structure type s that are predicted correctly, $s \in \{\alpha\text{-helix}, \beta\text{-sheet}, \text{coil}\}$. To measure the "quality" of the prediction on each type of secondary structure, Matthews' correlation coefficient was also used. For secondary structure type s,

$$C_s = \frac{(p_s \cdot n_s) - (u_s \cdot o_s)}{\sqrt{(p_s + u_s) \cdot (n_s + o_s) \cdot (p_s + u_s) \cdot (p_s + o_s)}},$$

where p_s is the number of positive cases that were correctly predicted; n_s is the number of negative cases that were correctly rejected; o_s is the number of overpredicted cases, and u_s is the number of underpredicted cases. These coefficients thus measure the differences of predictions for different types of structures.

Table 2

Number of residues, helix and sheet contents and names of protein sequences in each test group

Group	No. residue	Helix (%)	Sheet (%)	Proteins and their maximum homology with other proteins (%)	Average of maximum (%)
1	2417	29·6	21·2	1FC2-C(39·5), 2MT2(32·8), 1FXB(34·6), 2B5C(32·9), 3C2C(43·7), 2SNS(31·2), 2LHB(29·5), 1MBD(29·4), 1ETU(24·0), 2ACT(45·0), 1MCP-H(42·8), 1SBT(36·0), 2MDH-B(37·5), 3BCL(19·4)	34·2
2	2465	28·1	19·5	1GCN(44·8), 1PPT(38·9), 4PTI(34·5), 1UBQ(38·2), 3WRP(31·7), 156B(30·9), 2PAB-A(31·6), 1CY3(30·5), 4FXN(31·2), 2STV(26·6), 3RP2-A(31·2), 1TON(37·4), 2CAB(24·2), 4LDH(21·0), 2CTS(19·2)	31·5
3	2550	27·5	22·7	1INS-D(40·0), 2CI2-I(36·9), 1SN3(32·3), 1PCY(31·3), 1HMQ-A(29·2), 2AZA-B(30·2), 2HHB-B(41·1), 1LH1(30·1), 3GAP-A(24·5), 1FB4-H(41·5), 4PTP(41·9), 2CYP(21·8), 2APR(39·1), 3GRS(18·4)	32·7
4	2450	25·1	20·7	1MLT-A(46·2), 2OVO(33·9), 351C(35·4), 3FXC(30·6), 1CCR(34·2), 1PAZ(32·5), 1ECD(28·7), 2LZM(26·2), 3WGA-B(24·6), 1MCP-L(42·3), 2CGA-A(40·0), 1ABP(22·5), 2TBV-C(22·7), 1PHH(21·1)	31·5
5	2492	26·0	20·1	1CRN(37·0), 1NXB(37·1), 1CTF(39·7), 1RNT(29·8), 2CDV(29·9), 1RN3(27·4), 1PP2-L(38·3), 4ATC-B(30·1), 1GP1-A(25·5), 4SBV-C(27·0), 3EST(35·5), 5CPA(21·8), 4TLN(21·8), 3PGK(20·5)	30·1
6	2476	23·7	20·4	3RXN(38·5), 1FDX(37·0), 3ICB(36·0), 2GN5(32·2), 1CPV(35·2), 1LZ1(26·9), 2HHB-A(42·6), 2SOD-B(28·5), 1FC1-A(25·2), 9PAP(46·2), 1SGT(32·1), 4ATC-A(22·3), 1GDI-O(22·3), 4ADH(20·3)	31·8
7	2507	27·2	21·0	4CPA-I(35·1), 1CSE-I(34·9), 1CC5(33·7), 4FD1(31·1), 2SSI(32·7), 1BP2(41·5), 1FX1(27·9), 4DFR-B(27·0), 3ADK(28·9), 3PGM(26·1), 2CNA(24·9), 1RHD(23·5), 2APP(39·3), 2CPP(19·5)	30·4
8	2504	27·4	19·4	1INS-A(47·6), 1TGS-I(35·1), 2ABX-A(36·5), 1HIP(32·9), 1ACX(33·6), 2CCY-A(34·6), 155C(36·6), 4TNC(28·1), 1GCR(24·7), 1FB4-L(43·1), 1TIM-A(24·2), 2PRK(35·5), 2MDH-A(37·7), 7CAT-A(17·1)	33·4

The number of residues, helix and sheet contents, and the names of the protein sequences (subunits) in each test group. (1FC2-C, subunit C of 1FC2.) The number in the parenthesis after each protein name is the maximum homology between that sequence and all sequences in other groups (i.e. the training data set for that test group). The last column is the average of the maximum homology of each group.

(g) *A measure of statistical significance*

When comparing different prediction algorithms, we need to know whether the differences in prediction accuracy among them are statistically significant. Statistics theory gives us a method to compute the "significance interval" for the difference between 2 population proportions (Daniel, 1987). In the case of secondary structure prediction, the "proportion" is the percentage of residues in a set of test data whose secondary structure state has been correctly predicted. Assume the prediction accuracy of 2 algorithms are p_1 and p_2 for 2 test data sets of r_1 and r_2 residues, respectively, and the test data are randomly selected, then we say that we are $a \times 100\%$ confident that the accuracies of the 2 algorithms are really different if

$$|p_1 - p_2| > I,$$

where:

$$I = z(1+a/2) \cdot \sqrt{\frac{p_1(1-p_1)}{r_1} + \frac{p_2(1-p_2)}{r_2}}, \qquad (6)$$

z is the inverse cumulative normal distribution. For example, when $a = 0{\cdot}95$, $z(1+a/2) = 1{\cdot}96$; if $r_1 = r_2 = 20{,}000$, the significance interval is $I \approx 0{\cdot}9\%$; if $r_1 = r_2 = 4000$, $I \approx 2{\cdot}1\%$. If we choose $a = 0{\cdot}99$, $r_1 = r_2 = 20{,}000$, then $I \approx 1{\cdot}2\%$. Thus, the bigger the difference between 2 prediction accuracies, the more significant it is. For the same difference, the more test data used, the more significant it is (and the more confident we are). Equation (6) is used in this paper to determine whether the difference in the accuracies of 2 different predictions is statistically significant.

3. Experiments and Results

(a) *K-way cross-validation*

To evaluate the hybrid system, all the proteins in our database were randomly divided into eight groups. In each test, one group of proteins was used as the test data set and the rest as the training data set. The whole experiment consisted of eight such tests, i.e. eight independent runs of the hybrid system, each time on a different test data set. This way, there was no overlap between training data and test data, and every protein was used as test data once. This is the so-called "k-way cross-validation" testing procedure. Table 2 lists the proteins and the number of residues in each group, the α helix and β sheet contents in the group, as well as the degree of homology between proteins in different groups.

(b) *Window size and other choices*

Throughout this work, a window size of 13 residues was used. Each expert looked at 13 residues at a time and predicted the secondary structure state of the center residue in the window. The Combiner looked at the predictions of 13 residues from each expert and made a final prediction for the center residue. For each amino acid sequence in the test data set, the window was moved over the whole

sequence, and a prediction was made for every residue.

There were other choices that had to be made before starting our k-way cross-validation experiment with the hybrid system. They included (1) the number of hidden units and the number of training cycles for neural networks; (2) the threshold for "nearest neighbors" in MBR module; and (3) the coefficient C_f in the SM. If these choices were made according to the system's performance on the test data set, then they might be fine-tuned to fit the particular data set and make the system's accuracy appear higher than it really is. To avoid this, prior to the k-way validation experiment, a "pilot set" of 20 proteins was randomly chosen from the database, and the above choices were made based on the system's performance on this pilot set. (The pilot set consisted of: 1INS-A, 3RXN, 2MT2, 1CTF, 351C, 2CDV, 1HMQ-A, 1RN3, 1PP2-L, 4FXN, 2SOD-B, 1MBD, 1GP1-A, 1FB4-L, 4PTP, 1TON, 2PRK, 4ATC-A, 4LDH, 1PHH.)

(c) *MBR and SM: training and prediction*

In Memory-Based Reasoning module, first the distance matrices were computed using the training data set. There was one distance matrix for each position of the window, see equation (1) for details. Then for each segment (window) of the amino acid sequences in the test data set, $b_1, b_2, \ldots b_n$, the top 25 instances in the training data set that had the shortest distance to it were considered its neighbors. The strength of prediction (score) for each secondary structure state was the percentage of neighbors in that state weighted by the inverse of their distances. The structure state that had the highest score was taken as the prediction by MBR.

In the statistical module, the frequencies of singletons and pairs of amino acids within a window $a_1, \ldots, a_n$ were calculated for each structure state s_j in the training data set, to approximate the conditional probabilities $p(a_i|s_j)$s and $p(a_i, a_k|s_j)$s. Then for each segment of amino acid sequences in the test data set, $b_1, b_2, \ldots b_n$, these probability values were used to estimate the probability $p(s_j|b_1, b_2, \ldots b_n)$ according to equation (5) ($C_f = 1{\cdot}5$ was used in this work), where s_j is one of the secondary states (α-helix, β-strand and coil). The value of $p(s_j|b_1, b_2, \ldots b_n)$ was taken as the score of prediction for structural state s_j, and the state that had the highest score was taken as the prediction by SM.

(d) *Training neural networks*

One important issue in training neural networks by the Back-propagation algorithm is deciding when to stop training. If a network is trained through too many cycles, the network tends to memorize the training examples but generalizes poorly on the inputs that it has not been trained on (i.e. test data). One practice is to monitor the performance of the network being trained on the test data, and to stop training when the performance peaks. This strategy cannot be used in real situations where the true answer is truly unknown. We used the following techniques to solve this problem: (1) limiting the number of training cycles; (2) limiting the number of hidden units, thus the number of free variables (the "memory capacity") in the network; (3) when available, using a separate control data set to control when to stop training the network, that is, to monitor the performance of the network being trained on the control data set and stop training when the performance peaks.

A one-hidden-layer neural network was used as one of the three experts. This network is referred to as EXPERT-NN in the following discussion. A total of 21 input units was used to encode one residue, one unit for each of the 20 amino acid types plus one end marker. With a window size of 13 residues, there were $21 \times 13 = 273$ input units total. EXPERT-NN had three output units, one for each of the three secondary structure states (α-helix, β-sheet and coil). The network had only two hidden units. EXPERT-NN was trained up to 200 epoch cycles on the training data set, and the network weights that gave the best performance on the *training set* during training were saved as the final result of training. The activation of the output units were used as the score of prediction for the corresponding secondary structure.

The Combiner of our hybrid system was also a one-hidden-layer neural network. The Combiner took the outputs of the three experts as inputs and made final predictions based on these outputs. For every residue, each expert generated three numbers representing the prediction score for α-helix, β-sheet and coil, respectively. The Combiner took the predictions of 13 residues from each of the three experts as its input, thus it had $13 \times 3 \times 3 = 117$ input units. It also had three output units, one for each of the secondary structure states. As discussed in Methods and Materials, in order to train the Combiner, the training data set was divided into two halves, which will be referred to as $\{H_1\}$ and $\{H_2\}$ in the following discussion. The three experts were first trained on the first half of the data set $\{H_1\}$. Then they were applied on the second half $\{H_2\}$. Their outputs on $\{H_2\}$, $\{\text{Output}(H_2)\}$, were then used as input patterns for training the Combiner. Similarly, the three experts were also trained on $\{H_2\}$ and their outputs on $\{H_1\}$, $\{\text{Output}(H_1)\}$, were recorded. Finally, the Combiner was trained up to 200 epoch cycles, using $\{\text{Output}(H_2)\}$ as training data and $\{\text{Output}(H_1)\}$ as control data. The weights that gave the best performance on both $\{\text{Output}(H_1)\}$ and $\{\text{Output}(H_2)\}$ during training were saved as the result of training the Combiner. A total of 30 hidden units was used in the Combiner. Since there was a control data set here, the number of hidden units was less crucial here than in EXPERT-NN.

(e) *The hybrid system improved prediction accuracy*

Table 3 shows the results for the eight test data sets in our k-way cross-validation experiment. Table

Table 3
Prediction accuracy on test data sets

Group	No. sequence	No. residue	EXPERT-NN (%)	SM (%)	MBR (%)	Hybrid (%)
1	14	2417	60·7	62·8	64·4	65·3
2	15	2465	63·8	63·3	63·9	66·3
3	14	2550	62·2	63·6	64·7	66·2
4	14	2450	62·3	62·9	64·0	66·2
5	14	2492	63·2	62·4	64·4	66·6
6	14	2476	65·2	64·1	65·8	68·1
7	14	2507	62·3	63·8	63·1	65·1
8	14	2504	64·9	65·5	65·5	67·5
Total	113	19,861	63·1	63·5	64·5	66·4

The prediction accuracy on each test data set by the 3 experts and by the hybrid system. No. sequence is the number of sequences (subunits) in each group; No. residue is the number of residues in each group.

4 shows the accuracy for each sequence. Overall, for the prediction of secondary structures α-helix, β-sheet and coil, EXPERT-NN was 63·1% accurate, MBR was 64·5% and SM was 63·5%. The hybrid system was 66·4% accurate. The total number of residues used in the experiment was 19,861. According to the statistical significance measures described in equation (6), the improvement of the hybrid system over each expert was statistically significant (with higher than 0·99 confidence level). Thus we are highly confident that our hybrid system really improved the prediction accuracy.

The Matthews' correlation coefficients for each expert and for the hybrid system are shown in Table 5. All three experts had similar coefficients and produced better prediction on α helix and coil than on β strand. One reason for this might be that a single β strand can hardly be stable; more than one strand get stabilized when they interact with one another to form a β sheet; this interaction is often not local along the sequence and thus cannot be captured very well by the local approach. Thus, no matter what algorithm is used, β strand would still be the most difficult state to predict. The hybrid system improved the prediction for all the structure states.

(f) *A single small test data set is dangerous*

From Table 3 we computed the average difference in prediction accuracy among the three different experts for the same sets of test data, which was 0·9%. This shows that the overall accuracies of the three experts were very close. We also computed the average difference for the same expert on the eight different test data sets, which was 1·3%. Thus, if each test data set is observed independently, the difference in prediction accuracy caused by the different test data sets were at least as large as the difference brought about by the different experts. This observation argues strongly against using a single small test data set: (1) "statistical noise" can

Table 4
The accuracy on each protein sequence (subunit) by the three experts and the hybrid system

Protein	SM (%)	MBR (%)	EXPERT-NN (%)	Hybrid (%)
155C	64·9	74·6	64·9	70·9
156B	62·7	61·8	70·0	64·5
1ABP	59·8	53·3	57·8	57·5
1ACX	70·1	63·6	60·7	61·7
1BP2	52·0	55·3	52·0	52·8
1CC5	75·9	72·3	69·9	77·1
1CCR	67·6	70·3	70·3	73·0
1CPV	63·9	55·6	60·2	66·7
1CRN	50·0	56·5	50·0	52·2
1CSE-I	65·1	68·3	63·5	71·4
1CTF	55·9	57·4	50·0	54·4
1CY3	68·6	70·3	75·4	74·6
1ECD	44·9	43·4	36·8	45·6
1ETU	68·9	71·4	70·9	77·0
1FB4-H	71·6	75·5	65·9	71·2
1FB4-L	66·2	70·8	61·1	68·1
1FC1-A	56·8	60·2	60·2	68·3
1FC2-C	74·4	76·7	60·5	79·1
1FDX	72·2	79·6	70·4	72·2
1FX1	52·4	57·8	57·1	67·8
1FXB	82·7	75·3	70·4	77·8
1GCN	55·2	41·4	48·3	58·3
1GCR	45·4	52·9	56·9	62·9
1GD1-O	57·1	60·1	64·3	63·1
1GP1-A	64·7	60·3	65·8	65·2
1HIP	64·7	67·1	60·0	60·0
1HMQ-A	69·9	56·6	58·4	63·7
1INS-A	47·6	38·1	47·6	42·9
1INS-D	76·7	73·3	76·7	83·3
1LH1	70·6	61·4	68·0	72·5
1LZ1	73·8	66·9	70·0	68·5
1MBD	67·3	67·3	60·1	68·0
1MCP-H	64·4	75·7	61·3	81·5
1MCP-L	65·0	70·0	61·4	69·1
1MLT-A	42·3	50·0	50·0	46·2
1NXB	67·7	61·3	66·1	62·9
1PAZ	59·2	65·8	63·3	66·7
1PCY	58·6	64·6	65·7	67·7
1PHH	60·4	58·1	66·0	64·2
1PP2-L	65·4	70·7	66·9	69·2
1PPT	77·8	83·3	75·0	88·9
1RHD	64·2	66·2	64·8	66·2
1RN3	54·8	62·9	58·1	66·9
1RNT	64·4	61·5	68·3	67·3

Table 4 *(continued)*

Protein	SM (%)	MBR (%)	EXPERT-NN (%)	Hybrid (%)
1SBT	63·3	68·4	66·5	67·3
1SGT	66·7	75·4	65·0	78·3
1SN3	73·8	76·9	70·8	72·3
1TGS I	63·2	57·9	66·7	56·1
1TIM A	70·6	66·1	69·0	73·8
1TON	70·2	78·2	69·7	75·6
1UBQ	61·8	55·3	65·8	63·2
2ABX-A	89·2	78·4	81·1	81·1
2ACT	68·3	72·0	67·0	73·4
2APF	61·9	57·9	55·7	59·4
2APF	69·2	68·9	66·8	68·3
2AZA B	45·7	51·2	48·1	48·1
2B5C	65·9	67·1	63·5	55·3
2CAB	64·1	69·5	71·1	69·5
2CCY-A	75·6	63·8	79·5	83·5
2CDV	72·9	73·8	71·0	76·6
2CGA-A	64·9	74·7	58·4	72·2
2CI2	60·0	70·8	64·6	70·8
2CNA	57·8	58·6	60·3	59·1
2CPP	69·6	61·5	61·5	65·9
2CTS	68·0	62·5	64·1	69·1
2CYP	63·8	63·5	60·1	64·8
2GN5	69·0	65·5	62·1	70·1
2HHB-A	72·3	70·2	66·0	78·0
2HHB-B	61·0	59·6	47·9	61·6
2LHB	68·5	65·8	60·4	67·1
2LZM	60·4	64·6	61·6	61·6
2MDH-A	55·2	57·4	56·8	61·7
2MDH-B	48·0	53·2	47·1	52·0
2MT2	95·1	91·8	95·1	96·7
2OVO	62·5	64·3	60·7	66·1
2PAB-A	50·0	50·9	59·6	58·8
2PRK	63·8	69·9	63·8	71·3
2SNS	61·0	60·3	61·7	63·8
2SOD-B	66·2	67·5	72·2	71·5
2SSI	74·8	69·2	72·0	78·5
2STV	51·1	53·8	51·1	53·8
2TBV-C	59·5	62·0	60·1	64·5
351C	81·7	76·8	79·3	86·6
3ADK	64·4	68·0	61·9	69·6
3BCI	49·7	40·7	50·0	44·7
3C2C	71·4	82·1	59·8	68·7
3EST	70·1	77·7	64·9	79·3
3FXC	72·4	75·5	65·3	77·6
3GAT-A	50·5	60·1	54·8	56·7
3GRS	58·8	55·7	62·7	63·3
3ICB	85·3	89·3	86·7	90·7
3PGH	66·0	66·0	67·5	67·7
3PGM	63·9	67·8	67·4	68·3
3RP2-A	54·9	66·7	55·3	62·0
3RXN	82·7	84·6	84·6	84·6
3WGA-B	80·1	77·2	80·7	80·7
3WRP	75·2	70·3	66·3	73·3
4ADH	57·2	54·3	59·4	57·5
4ATC-A	58·4	61·3	61·0	63·5
4ATC-B	62·1	61·4	64·1	62·7
4CPA-I	78·4	67·6	73·0	70·3
4DFR-B	59·1	58·5	59·7	62·9
4FDP	67·9	73·6	75·5	72·6
4FXN	68·1	63·8	64·5	68·8
4LDH	59·8	58·3	59·2	61·3
4PTI	70·7	60·3	70·7	62·1
4PTI	71·4	82·5	68·8	77·8
4SBV-C	53·2	54·1	57·2	55·9
4TLN	57·9	62·3	58·2	65·8
4TNC	83·7	78·1	77·5	79·4
5CPA	60·9	63·8	63·8	66·4
7CAT-A	65·7	64·5	65·5	64·9
9PAP	70·3	80·7	70·3	76·4
Total	63·5	64·5	63·1	66·4

Table 5

The Matthews' correlation coefficients for each expert and the hybrid system on each structural state

Method	C_{coil}	C_α	C_β
SM	0·390	0·418	0·350
MBR	0·396	0·416	0·357
EXPERT-NN	0·395	0·383	0·333
Hybrid	0·429	0·470	0·387

Table 6

The percentage of total residues for which two experts produced the same secondary structure prediction

EXPERT-NN	MBR	SM	Hybrid
EXPERT-NN	76·6%	84·3%	82·9%
MBR		77·7%	82·0%
SM			83·6%

Table 7

Percentage accuracy

One correct	Two correct	Three correct	Three incorrect
76·6%	64·0%	50·6%	19·4%

make the same algorithm have different accuracies on different test data sets if the sets are small; (2) the difference among different algorithms, even if it truly exists, can be easily "buried" by such statistical noise. Thus, large or multiple test data sets should be used whenever possible.

(g) *Different algorithms made similar predictions*

The three experts used in our experiments did not only have similar overall prediction accuracies, but also made similar predictions for each sequence. Table 6 shows the percentage of the residues in the test data sets for which different experts produced the same predictions. On average, each pair of experts agreed with each other on about 80% of the total 19,861 residues. All three experts produced the same prediction on about 70% of the total residues (not shown in the Table). Table 7 shows the percentage of residues for which at least one expert was correct, at least two experts were correct, all three experts were correct and all three experts gave the same but wrong predictions. For about 20% of the residues, all three experts produced the same but wrong predictions. This, together with the information from Table 6 indicates that the "local rules" (the rules mapping short segments of amino acid sequences to secondary structures) obtained by the three very different experts were actually quite similar, but they did not apply quite as well to the test data. This may suggest an upper bound on the secondary structure prediction accuracy based on local information from the currently available data. In places where all algorithms were the same but

Table 8
Accuracy of predictions

Method	α Helix				β Sheet			
	Correct	Over	Under	Coef.	Correct	Over	Under	Coef.
SM	345	195	171	0·392	449	299	375	0·283
MBR	309	182	207	0·341	404	281	420	0·244
EXPERT-NN	314	181	202	0·351	421	316	403	0·238
Hybrid	353	162	163	0·445	450	234	374	0·335

The number of correct predictions (Correct), underpredictions (Under), overpredictions (Over) for α helix and β sheet by each expert and the hybrid system. Coef. is the Matthews' correlation coefficient (see Methods and Materials).

incorrect, the structures might be determined by non-local interactions. Among the residues where all three experts did agree with one another, they were correct for 71% of the residues. Thus, if we only consider the cases where all three experts agreed, we have a much higher prediction accuracy.

(h) *Homology between training and test data set*

It is known that if the training data and the test data are identical or highly homologous, then the prediction accuracy could be misleadingly high. However, when the degree of homology between training and test data was below 50%, we did not find strong positive correlations between the prediction accuracy and the degree of homology. For example, the degree of homology between 1GCN, 1MLT-A and 1INS-A and their training data were 44·8%, 46·2% and 47·6% respectively, and their prediction accuracies were quite low (see Table 4); whereas 2CTS, 3GRS and 7CAT-A had very low homology with their training data (19·2%, 18·4% and 17·1%, respectively), but their prediction accuracies were much higher.

(i) *Secondary structures as individual units*

Often it is more important to predict correctly the occurrence or absence of a secondary structure (α helix or β strand) as a whole rather than just to predict the states of individual residues. Thus the following criteria were also used in this work to evaluate the predictions of different methods: we took an α helix or β strand as an individual unit, and checked how many of these secondary structures were correctly predicted (positive cases), how many of them were not predicted at all (underpredicted), how many were predicted which do not exist in the real structures (overpredicted). Then a Matthews' correlation coefficient is calculated for each method. We found that the hybrid system had the most positive cases and the fewest overpredictions and underpredictions. (Note that this is in terms of number of secondary structures, not residues.)

Specifically, in this work an α helix is said to have been predicted if at least four continuous residues in a sequence are predicted to be in H state; a β strand is said to have been predicted if at least two continuous residues were predicted to be in E state. If the overlapping region between a real secondary structure and a predicted secondary structure of the same type is greater than half of the length of the real structure or the predicted structure, then the real secondary structure is considered to have been correctly predicted. If more than one predicted secondary structure overlaps with one real secondary structure, only one of the predicted secondary structures is considered as a correct prediction, and the rest are counted as overpredictions. If one predicted secondary structure overlaps with more than one real secondary structures is considered as correctly predicted, and the rest are counted as underpredictions. Table 8 lists the correct predictions, overpredictions, underpredictions and Matthews' coefficient for α helix and β strand by each expert and the hybrid system according to these criteria. (In calculating Matthews' coefficients, the residues between 2 helices (sheets) are considered to form 1 non-helix (non-sheet). The hybrid system produced the best result by this criteria as well.

No doubt the above criteria are not perfect. And the details such as the numbers 2 for β strand and 4 for α helix are to some extent arbitrary. However, we need some criteria to capture the intuitive notion of "how many secondary structures are predicted correctly". We believe the above criteria serves as an unbiased, first-order approximation to that. It provides a new perspective to evaluate different prediction methods. For example, SM is better than MBR and EXPERT-NN by this criteria, whereas that is not the case if we count the number of correctly predicted residue states (see Table 5).

(j) *An example*

Figure 4 shows the prediction for protein 1PAZ by each expert and the hybrid system. It illustrates the points discussed in previous sections. Note that the inputs from each expert to the Combiner in our hybrid system are the three prediction scores for each of the three states (α helix, β sheet and coil), not just the predicted states themselves; and the Combiner looks at the prediction scores of 13 posi-

```
          0         10        20        30        40        50        60
Protein:  ENIEVHMLNKGAEGAMVFEPAYIKANPQDTVTFIPVDKGHNVESIKDMIPEGAEKFKSKINENYVLTVTQ
Structure:-eeeeeeeeee--eeeeee--eeee----eeeeee-------ee-------------------eeee---

SM:       ----ehhhh-----hhhh-hh-ee------eeeee----------hh-h--hhhhhhhh----eeee--
MBR:      ---eehhh-h-----ee-----------eee-------------h--h--hhhhh-h----------
Expert-NN:---hhhhh------hehh----------eee--------------------hhh------eeeee--

Hybrid:   ----eee--------ee-----------eeee-------------------hhhhhh-----eee--

          70        80        90        100       110
Protein:  PGAYLVKCTPHYAMGMIALIAVGDSPANLDQIVSAKKPKIVQERLEKVIA
Structure:-eeeeee---------eeeeee------hhhhhh----hhhhhhhhhhh-

SM:       ----eee-----hhhhhhhee-------hhhhhh----hhhhhhhh----
MBR:      ----ee----------eeeee-------hhhhh-----hh-hhhhhhh--
Expert-NN:---eee------hhhhheeee---------hhh-------hhhhhhhh--

Hybrid:   ----ee------hhheeeeee--------hhhh-----hhhhhhhhhh--
```

Figure 4. The secondary structure prediction generated by SM, MBR, EXPERT-NN and the hybrid system. Structure indicates the secondary structure assignment by the DSSP program.

tions at a time. That is why we can see that in certain cases the Combiner can override the majority of the three experts, such as between residue 0 and 10 of 1PAZ. In some places, all three experts made the same but wrong predictions. For example, there is a short β strand between residue 40 and 50 that none of the experts predicted; and they all predicted a helix between residue 50 and 60 that does not exist in the real structure. In both cases the Combiner made the same mistake also. None of the experts could always make better predictions than others. For example, SM is the only one that predicted the sheet between residue 20 and 30. MBR is the only one that did not give the false prediction of a helix between residue 80 and 90, and EXPERT-NN made fewest mistakes between residue 50 and 70.

(k) *Comparison with other methods*

Qian & Sejnowski (1988) used a "cascaded neural network" system in secondary structure prediction and achieved 64·3% accuracy on a test set of 15 proteins (containing 3520 residues). Their system contained two networks: the first network took amino acid sequences as inputs and produced the initial prediction; the second network "cleaned up" this initial prediction to produce final predictions. This system could also be seen as a hybrid system but with only one expert. We applied their method to our eight test data sets. Table 9 shows the results. This was done not only to compare the final results, but also to see whether adding two more experts could really help. The overall prediction accuracy of the cascaded system on our test data sets was 64·0%, which, on a much larger scale (19,861 *versus* 3520 residues), confirmed Qian & Sejnowski's results. However, the improvement of the cascaded network over a single network was only 0·5%, not 1·5% as reported in their paper. According to our statistical significance measure (equation (6)), both 0·5% for 19,861 residues and 1·5% for 3520 residues were not statistically significant differences at confidence level 0·95. We also noticed that there was

Table 9

The accuracy on the eight test data sets by Cascaded networks of Qian & Sejnowski (1988)

Group	No. sequence	No. residue	Single network (%)	Cascaded network (%)
1	14	2417	61·9	62·5
2	15	2465	64·3	64·3
3	14	2550	62·5	63·2
4	14	2450	62·7	62·9
5	14	2492	63·3	64·3
6	14	2476	65·5	66·6
7	14	2507	62·6	62·9
8	14	2504	65·3	65·5
Total	113	19,861	63·5	64·0

some difference in prediction accuracy (0·4%) between their single network and our EXPERT-NN, even though they were both trained and tested on the same data sets. The reason was that according to Qian & Sejnowski's method, the performance of their network on the test data set was monitored during training. The network weights that performed the best on the test set were saved and used. Whereas in our work, the EXPERT-NN never saw the test data set during training (see Methods and Materials).

The GOR III algorithm by Gibrat *et al.* (1987) was reported to have achieved 63% prediction accuracy by using correlations between certain pairs of amino acids and secondary structures. Biou *et al.* (1988) further improved the GOR III algorithm by combining its result with that of two other algorithms, the Homologue method and the bit pattern method, achieving a reported accuracy of 65·5% (we refer to this combined algorithm as GOR-Combined in the following discussion). We ran the GOR-Combined program on protein sequences in our database. Since their program contained the statistics calculated using their database, i.e. their training data, we divided our database into two groups. Group A contained sequences that were identical or more than 50% homologous to their training data. Group B contained the rest of the sequences. There were 64 sequences in group A and 49 sequences in group B. Apparently group B should be used as the test data to compare the GOR-Combined against other algorithms, because a prediction algorithm could easily have a very high prediction accuracy on protein sequences that are either identical or highly homologous to its training data, which cannot be used as an objective assessment of the algorithm's prediction accuracy. For group B, the GOR-Combined was 62·4% accurate. This is 3% lower than their reported result. One reason for this might be that GOR-combined algorithm used certain rules to combine the outputs of different methods, and those rules did not work quite as well for proteins not in its database. We used the 64 protein sequences in group A to train our hybrid system and applied it to the 49 protein sequences in group B. It was 65·3% accurate. This

Table 10

Accuracies of different algorithms for three states (helix, sheet, coil) prediction

Method	Accuracy (%)
Lim (1974)	59
Chou & Fasman (1974)	50
Levin *et al.* (1986)	62·2
GOR III	63
Qian & Sejnowski (1988)	64·3
Holley & Karplus (1989)	63·2
Hybrid	66·4

is about 1% lower than the average accuracy of the hybrid system in the k-way cross-validation experiment. We believe this was due to the smaller training set used here, which had only 64 protein sequences.

Table 10 lists the results of several other algorithms. The results were obtained from each author's original report except those by Lim (1974) and Chou & Fasman (1974), because in their original reports they used the same data set for both training and testing. Kabsch & Sander (1983*b*) assessed the accuracies of these two algorithms with separate test data, and the results were included in the Table instead. Among these, our hybrid system was tested with the largest set of protein data and it gave the highest prediction accuracy.

4. Discussion

The idea of combining the strength of different methods is not entirely new in either machine learning research (Wolpert, 1990) or protein secondary structure prediction. For example, Biou *et al.* (1988) used certain rules to combine three methods. However, the authors did not explain how their rules were generated in the first place. Thus it is difficult for us to justify the use of those rules. In our hybrid system, the Combiner learns how to combine the outputs of different experts automatically from the training data. A novel procedure has to be developed to train the Combiner because different experts can have very different behaviors. For example, after training, some experts can be 100% correct on the training data set while others may be only 70% correct on the training data even though they have very similar prediction accuracies for proteins not in the training set. Our training procedure for the Combiner can cope with experts that have such different characteristics.

This work showed that although different algorithms may have very similar overall secondary structure prediction accuracies, their detailed predictions can be different. No single algorithm always gives a better prediction than others. A combination of them can produce a statistically significant improvement over each individual method. We developed a way to train a Combiner, which learned to combine the outputs of different experts automatically. A neural network was used

as the Combiner in this work. But it is not the only choice. A MBR system, for example, can also be used as a Combiner. This paper is the first place where the SM algorithm and the particular MBR distance function have been introduced. Their accuracy were as good as or even better than any other single algorithm reported to date for secondary structure prediction. They deserve a more detailed discussion, which is beyond the scope of this paper and is done elsewhere (X. Zhang, unpublished results). The techniques we used to control the training of artificial neural networks were not only objective but also effective. For a single one-hidden-layer network, the accuracy was 63·1% with our techniques (to control training purely based on the training data). Whereas the other approach, to monitor the performance of the network on the test data during training, was 63·5%. The difference between them was only 0·4%. Thus our techniques produced near-optimal training.

One of the reviewers of this paper raised the issue of whether residues assigned to state G by the DSSP program (Kabsch & Sander, 1983*a*) should be considered as in helix, especially when they are adjacent to state H. In our original experiments, we wanted to make our result directly comparable with results obtained by other researchers, such as Qian & Sejnowski (1988), since the main point of this paper is that for the same secondary structure assignment, the hybrid system gives better prediction than other algorithms. Thus we used the same assignment as Qian & Sejnowski (1988), i.e. only considering H for α helix and E for β strand. After we received the reviewer's comments, we did the following experiment: we assigned G states to be helix if they are adjacent to H, otherwise assigning them to be coil. This way, among the 19,861 residues in our database, 162 residues (0·8% of the total residues) were assigned differently, i.e. to helix instead of coil. Then we compared the original prediction of our hybrid system with this new assignment. It is 66·1% accurate. This is very close to the original accuracy of 66·4%. The change in accuracy (0·3%) is much smaller than the change in the assignment (0·8%). This means that even though the hybrid system was trained with a different assignment, it can still predict correctly most of the new assignment. This is in accordance with observations by other researchers (e.g. Richardson & Richardson, 1988) that there are certain ambiguities on secondary structure boundaries assigned by DSSP.

Good criteria for evaluating and comparing different prediction algorithms are crucial for the progress of this research field. In this work, we made use of the significance interval measure from statistics, which could tell us whether the differences observed are significant or not, and what factors can influence that. We emphasize the importance of the fact that in our tests, the hybrid system never looked at the test data during training, thus making the performance of the system on the test data as objective as possible. The k-way cross-validation allowed us to test our hybrid system with as many data as we have, and yet still avoided overlapping between the test data and training data. Some researchers have used one protein in each test group, thus maximizing the training data size. However, the extremely large amount of computation in our work prevented us from doing that (i.e. $k = 113$, the total number of protein sequences of our database). We choose $k = 8$, which did not reduce the size of each training data set very much, and yet cut the amount of computation dramatically. Even so, a large amount of computation was still needed to carry out our experiment. This involved (1) computing many statistics for SM and distance matrices for MBR; (2) pattern matching and sorting through the whole database to find neighbors in MBR; and (3) training many neural networks with large numbers of input/output examples. The experiment was done on a massively parallel computer Connection Machine CM-2. The particular machine we used had 4096 processors. In general, CM-2 can have up to 65,536 processors.

There are many important issues in protein secondary structure prediction, such as: (1) is "the percentage of correctly predicted residues" the best measure for success? (2) What is the best way to assign the secondary structures to a protein once its three-dimensional co-ordinates are known? (3) What is the right criteria for homology in selecting test/training data? A comprehensive discussion of these issues is beyond the scope of this paper. The emphasis here is to demonstrate that our hybrid system gives significantly better performance than individual algorithms and all previous methods, using the same criteria in selecting data and the same accuracy measure as used by other researchers.

We are grateful to Eric Lander and Tau-Mu Yi for valuable comments and suggestions on several drafts of this paper. We thank Christian Sander for providing us with the DSSP program and Anand V. Bodapati for helpful discussions. We also thank the anonymous reviewers who gave us insightful comments.

References

Bahadur, R. R. (1961). On classification based on responses to a dichotomous items. In *Studies in Item Analysis and Prediction,* chapt. 10, Stanford University Press.

Biou, V., Gibrat, J. F., Levin, J. M., Robson, B. & Garnier, J. (1988). Secondary structure prediction: combination of three different methods. *Protein Eng.* **2**, 185–191.

Chou, P. Y. & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry,* **13**, 222–244.

Daniel, W. W. (1987). *Biostatistics: A Foundation for Analysis in the Health Sciences.* 4th edit., John Wiley & Sons.

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.

Gibrat, J.-F., Garnier, J. & Robson, B. (1987). Further

developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* **198**, 425–443.

Holley, L. H. & Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 152–156.

Kabsch, W. & Sander, C. (1983*a*). Dictionary of protein secondary structures: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kabsch, W. & Sander, C. (1983*b*). How good are predictions of protein secondary structure? *FEBS Letters*, **155**, 179–182.

Kneller, D. G., Cohen, F. E. & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **214**, 171–182.

Lathrop, R. H., Webster, T. A. & Smith, T. F. (1987). nishiARIADNE: pattern-directed inference and hierarchical abstraction in protein structure recognition. *Commun. A.C.M.* **30**, 909–921.

Lazarsfeld, P. F. (1961). The algebra of dichotomous systems. In *Studies in Item Analysis and Prediction*, chapt. 8, Stanford University Press.

Levin, J. M., Robson, B. & Garnier, J. (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Letters*, **205**, 303–308.

Levitt, M. (1983). Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.* **170**, 723–764.

Lim, V. I. (1974). Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.* **88**, 873–894.

McClelland, J. L. & Rumelhart, D. E. (eds) (1986). *Parallel Distributed Processing*. MIT Press.

Nishikawa, K. & Ooi, T. (1986). Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochim. Biophys. Acta*, **871**, 45–54.

Qian, N. & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865–884.

Richardson, J. S. & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of α helices. *Science*, **240**, 1648–1652.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing*, MIT Press.

Sholnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science*, **250**, 1121–1125.

Stanfill, C. & Waltz, D. (1986). Toward memory-based reasoning. *Commun. A.C.M.* **29**, 1213–1228.

Taylor, W. R. & Thornton, J. M. (1984). Recognition of super-secondary structure in proteins. *J. Mol. Biol.* **173**, 487–514.

Wolpert, D. H. (1990). *Stacked Generalization*. Tech. Report LA-UR-90-3460, LANL.

Edited by F. Cohen

PROTEINS: Structure, Function, and Genetics 23:38–48 (1995)

Rigid Domains in Proteins: An Algorithmic Approach to Their Identification

William L. Nichols,[1] George D. Rose,[2] Lynn F. Ten Eyck,[1,3] and Bruno H. Zimm[1]
[1]*Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093;*
[2]*Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205;* [3]*San Diego Supercomputer Center, San Diego, California 92186-9784*

ABSTRACT **A rigid domain, defined here as a tertiary structure common to two or more different protein conformations, can be identified numerically from atomic coordinates by finding sets of residues, one in each conformation, such that the distance between any two residues within the set belonging to one conformation is the same as the distance between the two structurally equivalent residues within the set belonging to any other conformation. The distance between two residues is taken to be the distance between their respective α carbon atoms. With the methods of this paper we have found in the deoxy and oxy conformations of the human hemoglobin $\alpha_1\beta_1$ dimer a rigid domain closely related to that previously identified by Baldwin and Chothia (J. Mol. Biol. 129: 175–220, 1979). We provide two algorithms, both using the difference-distance matrix, with which to search for rigid domains directly from atomic coordinates. The first finds all rigid domains in a protein but has storage and processing demands that become prohibitively large with increasing protein size. The second, although not necessarily finding every rigid domain, is computationally tractable for proteins of any size. Because of its efficiency we are able to search protein conformations recursively for groups of non-intersecting domains. Different protein conformations, when aligned by superimposing their respective domain structures, can be examined for structural differences in regions complementing a rigid domain.**

Key words: difference-distance matrix, hemoglobin rigid core, structure search

INTRODUCTION

Structural domains in proteins have been defined in numerous ways, among the better known being visually recognizable conformational regions[1] and sets of proximate residues within difference maps.[2] More quantitative definitions include clustering,[3] use of cutting planes,[4] minimization of interfacial surface area,[5] maximization of solvent exclusion,[6] minimization of specific volume,[7] isolation of coherent regions from normal mode analysis,[8] and maximization of compactness.[9]

In this paper tertiary structures existing in different protein conformations define a rigid domain if the distance between any two residues of the rigid domain structure in one conformation is the same as the distance between the two equivalent residues of the rigid domain structure in every other conformation. The distance between two residues is defined to be the distance between their respective α carbon atoms, which can be found from atomic coordinates. The tertiary structures defining a rigid domain in different protein conformations are geometrically congruent and can be superimposed by aligning their equivalent residues. The residues of a domain (we shall refer to a rigid domain simply as a domain) do not have to be sequentially or spatially contiguous. The conformations being searched for domains must have their primary sequences at least partially aligned prior to implementing the algorithms of this paper. For this reason our methods are easily applied to the T and R states of an allosteric protein. No assertions are made about the persistence of structural rigidity of a domain along transitional pathways between conformations.

Figure 1 illustrates the concept of a domain in an eight-residue peptide with two conformations, A and B. The heavy line connecting successive α carbon atoms represents the peptide backbone. Distances between all pairs of α carbon atoms are shown by dashed lines in conformation A. Five residues form a domain within the peptide, as shown by the dashed lines in conformation B, which indicate that the distances between all pairs of residues in the domain are the same in both conformations. No one of the other three residues can be included in the domain because its distance from at least one of the five residues of the domain is not the same in conformation A as in conformation B.

We would like to have tertiary structures that are nearly but not exactly congruent to each other nev-

Received December 19, 1994; revision accepted February 9, 1995.
Address reprint requests to William L. Nichols, Department of Chemistry and Biochemistry 0317, University of California, San Diego, La Jolla, California 92093.

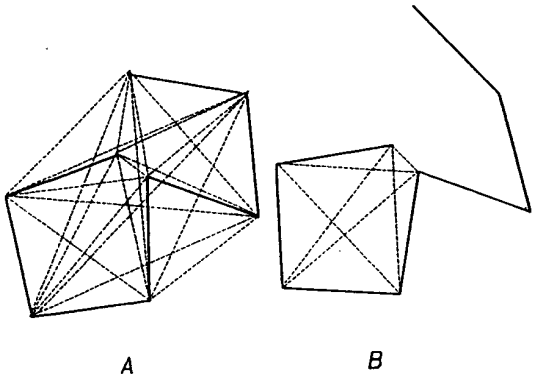

Fig. 1. A conformational change in an eight-residue peptide in which five residues form a domain. Dashed lines connect all pairs of residues in the initial structure (**A**) on the left, but only pairs from the five residues that form a domain in **B** on the right. The solid line represents the α-carbon backbone of the peptide.

ε=0.30Å

J	N(J)	C(R,J)
2	643	861
3	5341	11480
4	29121	111930
5	114643	850668
6	343572	5245786
7	808298	26978328
8	1520258	26978328
9	2311635	445891810
10	2861660	1471442973
11	2895704	4280561376
12	2398436	11058116888
13	1623937	25518731280
14	894886	52860229080
15	398055	98672427616
16	141005	166509721602
17	38937	254661927156
18	8098	353697121050
19	1196	446775310800
20	112	513791607420
21	5	538257874440

Largest rigid domains:

{7 8 10 20 21 23 24 25 26 27 28 29 30 33 35 36 37 38 39 41 42} *RMS* = 0.205Å
{7 8 10 20 21 23 24 25 26 27 28 29 30 33 36 37 38 39 40 41 42} *RMS* = 0.209Å
{7 8 10 20 21 23 24 25 26 27 28 29 30 31 33 35 36 37 38 39 41 42} *RMS* = 0.208Å
{7 8 20 21 23 24 25 26 27 28 29 30 31 33 36 37 38 39 40 41 42} *RMS* = 0.213Å
{7 8 20 21 23 24 25 26 27 28 29 30 31 32 33 36 37 38 39 40 41 42} *RMS* = 0.213Å

Residues common to all of the largest rigid domains:

{7 8 20 21 23 24 25 27 28 29 30 33 36 37 38 39 41 42}

Fig. 2. An exhaustive determination of domains within the N-terminus, A, B, and C helices of the α_1 monomer of human hemoglobin with $\varepsilon = 0.30$ Å. The number of residues in a domain is J. The number of domains with J residues is $N(J)$. The number of possible sets with J residues that can be found from all the $R = 42$ residues is $C(R,J)$. For the peptide of this example, the largest rigid domains have 21 residues. The residues in each of the five largest domains are listed at the bottom of the figure along with the 18 residues common to all. Root-mean-square values for the fit of the oxy onto the deoxy domains are given to the right of each.

ertheless to define a domain. For these structures, distances between equivalent residues will differ among conformations. Differences in distances can arise from insignificant dissimilarities between structures defining a domain or from experimental uncertainty in coordinates. To allow us to include geometrically incongruent structures as domains we generalize our definition of a rigid domain by specifying a parameter ε so that the distance between two residues of a domain in one conformation can differ from the distance between the structurally equivalent residues of the domain in another conformation by as much as ε. The number of residues included in a domain then depends on the value chosen for ε. Domains found with small values of ε reveal more detailed differences in structure between conformations, while domains found with larger values of ε identify gross similarities among conformations. When searching a group of protein conformations for domains, a good initial choice for ε is the precision measure of the atomic coordinates. The efficacy of the methods of this paper for identifying structural similarities in protein conformations is due in part to their not relying upon a least-square measure of similarity to identify domains but only upon the maximum absolute deviation in inter-residue distance as given by ε.

Two distinct domains may have residues in common or be entirely disjoint. The minimum number of residues in a domain can be as small as two and still be consistent with our definition, but the maximum number of residues in a domain is limited only by the number of residues in the protein. The hemoglobin molecule can serve to illustrate these points. Hemoglobin consists of two monomers, termed α and β. Two copies of each monomer associate to form the native tetramer. The hemoglobin structure has been solved by X-ray crystallography in both oxy and deoxy forms. Coordinates for human deoxyhemoglobin[10] and human oxyhemoglobin[11] were obtained from the Protein Data Bank[12] as entries 2HHB and 1HHO, respectively. The search of residues 1–42 of the α_1 monomer (the N-terminus, A, B, and C helices) for domains whose inter-residue distances differ by no more than 0.30 Å between deoxy and oxy conformations finds 643 domains with two residues each. The five largest domains have 21 residues each and have 18 residues in common. These results, appearing in Figure 2, will be discussed further in following sections.

Different protein conformations can be aligned by superimposing a common domain. A measure of how well domain structures align is the root-mean square (RMS) fit of superimposed domain residues:

$$\mathrm{RMS} = \sqrt{\frac{\sum_i \Delta x_i^2 + \Delta y_i^2 + \Delta z_i^2}{R}}.$$

$\Delta x^2 + \Delta y^2 + \Delta z^2$ is the squared distance between corresponding residues in two different superimposed domain structures of R residues each. RMS comparisons of entire conformations tend to conceal

small differences in structure, differences that are readily apparent when domain structures found with a sufficiently small ε are superimposed. Many others have addressed aspects of structural alignment, among them Vriend and Sander[13] and Holm and Sander.[14]

Within the hemoglobin tetramer, the identical $\alpha_1\beta_1$ and $\alpha_2\beta_2$ dimers undergo a quarternary reorientation relative to one another with the change from the deoxy to the oxy conformation.[15] Baldwin and Chothia[16] found a group of residues within the $\alpha_1\beta_1$ dimer interface whose α carbon atoms remain rigid during this allosteric transition. With our methods we find such a set of α carbon atoms as well and refer to it as the rigid core of the dimer. The dimer core can be used to align the oxy and deoxy structures to reveal conformational changes with ligand binding.

METHODOLOGY

Calculating the Difference-Distance Matrix

The initial computational step for finding domains from atomic data is to construct distance matrices and to use them to find the difference–distance matrix. For a given conformation, elements of a distance matrix D_{ij} are the distances between α carbon atoms i and j. With $D_{ij}^{(1)}$ the distance matrix for one conformation and $D_{ij}^{(2)}$ the distance matrix for another conformation, the difference–distance matrix[17–19] is given by the absolute value of the matrix difference,

$$\Delta_{ij} = |\, D_{ij}^{(1)} - D_{ij}^{(2)}|. \tag{1}$$

For computational efficiency, a matrix δ_{ij} is defined such that if

$$\begin{aligned} \Delta_{ij} > \varepsilon,\ \delta_{ij} &= 0, \\ \Delta_{ij} \leq \varepsilon,\ \delta_{ij} &= 1. \end{aligned} \tag{2}$$

Residue pairs i, j with a change of inter-residue distance (between α carbon atoms) of no more than ε Å have matrix elements $\delta_{ij} = 1$; otherwise $\delta_{ij} = 0$. The value chosen for ε will depend on the purpose of the calculation, being small if we seek subtle differences between conformations or large if we are searching for gross similarities among conformations.

Exhaustive Search For Domains

An exhaustive search for domains within a polypeptide constructs all rigid residue pairs and from this set finds rigid triples, quadruples, and so forth, until all combinatorial possibilities have been considered. That is, an exhaustive search begins with the set of all pairs of residues i, j for which $\delta_{ij} = 1$ and from this set finds the set of all distinct rigid triplets of residues i, j, k. A triplet is rigid if $\delta_{ij} = 1$ for i and j any of the residues in the triplet. The search is iteratively enlarged by finding every domain with $J + 1$ residues from each domain with J residues until all possible combinations of residues have been exhausted. Figure 2 illustrates this method. The number of residues in a domain is J while $N(J)$ is the number of distinct domains with J residues. The binomial coefficient $C(R,J)$ is the number of possible subsets of J residues, rigid or otherwise, that can be found in a set of R residues. The number of domains $N(J)$ for each J can be fairly large, although it is usually much smaller than the number of possible subsets $C(R,J)$. From Figure 2, for example, 4,280,561,376 subsets of 11 residues exist within a set of 42 residues, but only 2,895,704 domains of 11 residues each can be identified when $\varepsilon = 0.30$ Å in the first 42 residues (the N-terminus, A, B, and C helices) of the α_1 monomer of human hemoglobin. Figure 2 is further examined in the Discussion section. Searching for domains in this way is computationally demanding because the number of subsets in a set of R residues increases rapidly with R. We need to look for a faster way to find domains.

An Incomplete but Fast Search for Domains

We now introduce an alternative method for finding domains that is computationally feasible for a polypeptide of arbitrary length. With this method, only a single domain is identified for J residues [rather than the $N(J)$ domains required for an exhaustive search], with J near the maximal domain size. The complement of this domain is then searched exhaustively to identify all larger domains that include it as a subdomain. The method saves considerable computational effort and will still find domains suitable for conformational comparison.

We assert that a residue i differing by a small amount in relative position from one conformation to another will have many of its $\delta_{ij} = 1$, while a residue differing by a large amount in relative position will have many of its $\delta_{ij} = 0$. To quantify the changes in residue position, sums S_i of δ_{ij} are evaluated for each residue i over all other residues j in the polypeptide:

$$S_i = \sum_j \delta_{ij}. \tag{3}$$

The residues for which position differs the least between conformations tend to have the largest S_i while those for which position differs the most between conformations tend to have the smallest S_i.

The search for a domain is initiated by choosing an integer N_s and finding all those residues i for which $S_i \geq N_s$. The set of residues for which this condition is true is defined as $U_\varepsilon(N_s)$. $U_\varepsilon(N_s)$ will be the entire protein when N_s is zero and will have only the more rigid residues as N_s approaches the number of residues in the protein. $U_\varepsilon(N_s)$ is usually not itself a domain, because distance matrix elements between residues in $U_\varepsilon(N_s)$ for one conformation can differ by more than ε from those for other conformations.

However, the following strategy will find at least one domain in $U_\epsilon(N_s)$, if any exist.

For each residue i in $U_\epsilon(N_s)$ find all other residues j in $U_\epsilon(N_s)$ for which $\delta_{ij} = 0$, that is, such that the residue pair i, j is not rigid. The residue i that has the largest number N_M of residues j for which $\delta_{ij} = 0$ is removed from $U_\epsilon(N_s)$, leaving a subset of $U_\epsilon(N_s)$ with one less residue. [The subsets of $U_\epsilon(N_s)$ will be affected by the order in which residues with the same value for N_M are deleted.] The reduction is repeated until $N_M \leq 1$. This subset of $U_\epsilon(N_s)$ is a domain except for possible non-rigid pairs of residues. A domain $D_\epsilon(N_s)$ can then be constructed from this subset by removing all non-rigid pairs.

$D_\epsilon(N_s)$ can be enlarged by searching its complement exhaustively to find all domains that preserve $D_\epsilon(N_s)$ as a subdomain. Among these will be domains found by adding back some residues that were previously removed as non-rigid pairs in $U_\epsilon(N_s)$, but other domains are often discovered as well.

Constructing a domain of J residues using the method described in this section is much faster than finding one by exhaustive enumeration of all $N(K)$ possibilities, as K grows from 2 to J. By choosing N_s appropriately, the domains found by enlarging $D_\epsilon(N_s)$ will generally be maximal or, if not, will have residues in common with the maximal domains. The algorithm to find $D_\epsilon(N_s)$ is most efficient for values of N_s near R, the number of residues in the protein, because construction of the domain $D_\epsilon(N_s)$ is computationally fast, and the exhaustive search through the complement of $D_\epsilon(N_s)$ will not have to find many larger domains.

A domain can be used to align protein conformations by translating the centroid of the domain for each conformation to the coordinate origin and rotating the domain of one conformation onto that of the others with methods originally described by Kabsch.[20,21] The resulting transformed coordinates give a least-square fit between the domains of the different conformations. The entire protein can now be visually or numerically investigated for conformational differences in other regions.

A summary of the above algorithm follows.

I. Read the coordinates of all residues i for each conformation.
II. Construct the distance and difference-distance matrices.
  A. Choose ϵ. [See the remarks about choosing ϵ after Eq. (2) above.]
  B. Calculate the difference-distance matrix Δ_{ij} for all pairs of residues i, j.
  C. If $\Delta_{ij} > \epsilon$ then $\delta_{ij} = 0$; otherwise $\delta_{ij} = 1$.
III. Find a domain (not necessarily the largest).
  A. Choose N_s.
  B. Calculate S_i for each residue i.
  C. For each i, if $S_i \geq N_s$ then include residue i in the set $U_\epsilon(N_s)$.
  D. For each i in the set $U_\epsilon(N_s)$, find all residues j also in $U_\epsilon(N_s)$ for which $\delta_{ij} = 0$.
  E. Remove from $U_\epsilon(N_s)$ that residue i that has the most other residues j for which $\delta_{ij} = 0$.
  F. When for every residue i remaining in $U_\epsilon(N_s)$ at most only one other residue j can be found for which $\delta_{ij} = 0$, remove both i and j from $U_\epsilon(N_s)$ to give $D_\epsilon(N_s)$. Otherwise repeat III.D. and III.E.
IV. Search for larger domains.
  A. Examine each residue j in the complement of $D_\epsilon(N_s)$ to see if $\delta_{ij} = 1$ for all residues i in $D_\epsilon(N_s)$.
  B. For each j for which all $\delta_{ij} = 1$ in IV.A., include j in $D_\epsilon(N_s)$ to form a domain one residue larger.
  C. Repeat IV.A. and IV.B. with each such domain until no larger domains can be found.

DISCUSSION

We illustrate the methods defined above by searching for domains in the first $R = 42$ residues of the α_1 monomer of human hemoglobin (the N-terminus, A, B, and C helices). (The reader will please note that we are using hemoglobin as a convenient example for the application of these methods; we make no pretense to a thorough study of this protein in this paper.) As previously, J is the number of residues in a domain, $N(J)$ is the number of domains with J residues, and $C(R,J)$ is the number of possible subsets of J residues in a set of R residues. Figure 2 outlines an exhaustive search for domains within this peptide when $\epsilon = 0.30$ Å. The number of rigid residue pairs is 643 while the number of possible pairs of residues is 861. Similarly, the number of possible triplets is 11,480, but only 5,341 of them are rigid. The number of possible sets $C(R,J)$ grows combinatorially with J, while the number of domains eventually converges to 5 when J is 21. The 18 residues common to all five largest domains are listed at the bottom of the figure. The rigidity of these five largest domains is assessed by superimposing the deoxy and oxy structures,[20,21] with RMS fits as shown in the Figure. For conformation alignment all the largest domains are effectively the same, as can be seen in Figure 3.

The computational demands of an exhaustive search for domains are apparent in Figure 2. The number of sets with J or fewer residues that could be rigid is the sum of all the binomial coefficients $C(R,J)$ from 2 through J, a number that grows exponentially with J. The fast search avoids such an encumbrance by finding only one of the $N(J)$ domains, the domain $D_{0.30}(N_s)$, and exhaustively enlarging only this one domain. With a value $N_s = 25$, the search of the difference-distance matrix results in a set $U_{0.30}(25)$ of 35 residues:

$U_{0.30}(25) =$
{1 2 3 6 7 8 9 10 13 14 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42}.

The residue belonging to the most non-rigid residue pairs is residue 14, which is not rigid with $N_M = 12$ other residues in $U_{0.30}(25)$ (see table below). After deleting residue 14 from $U_{0.30}(25)$, the subsequent search for the residue with the largest sum N_M finds residue 22, with ten non-rigid pairs. Iteration until N_M is not larger than 1 results in the removal of 13 residues from $U_{0.30}(25)$, which leaves a set of 22 residues, the last one removed being residue 40. The following summarizes the deletion of residues from $U_{0.30}(25)$:

J	EXCLUDED RESIDUE	N_M
35	14	12
34	22	10
33	34	8
32	26	7
31	23	6
30	10	6
29	21	5
28	7	5
27	25	4
26	29	3
25	8	3
24	42	2
23	40	2

A set of 22 residues with two non-rigid residue pairs, (19,41) and (32,35), remains after the removal of residue 40. Deleting both non-rigid residue pairs leaves a domain $D_{0.30}(25)$ with 18 residues:

$$D_{0.30}(25) = \{1\ 2\ 3\ 6\ 9\ 13\ 18\ 20\ 24\ 27\ 28\ 30\ 31\ 33\ 36\ 37\ 38\ 39\}.$$

$D_{0.30}(25)$ is only one of the 8,098 domains with 18 residues found exhaustively in Figure 2. An exhaustive search through the complement of $D_{0.30}(25)$ finds four domains with 20 residues each, the two additional residues being one from each of the non-rigid residue pairs (19,41) and (32,35):

{1 2 3 6 9 13 18 19 20 24 27 28 30 31 32 33 36 37 38 39} RMS = 0.191 Å
{1 2 3 6 9 13 18 19 20 24 27 28 30 31 33 35 36 37 38 39} RMS = 0.186 Å
{1 2 3 6 9 13 18 20 24 27 28 30 31 32 33 36 37 38 39 41} RMS = 0.213 Å
{1 2 3 6 9 13 18 20 24 27 28 30 31 33 35 36 37 38 39 41} RMS = 0.205 Å

The RMS value for the superposition of each oxy domain upon its deoxy counterpart is listed after each domain. No other residues could be found in the complement of $D_{0.30}(25)$ that would fit rigidly in any of the four domains listed above.

We now compare the above with the exhaustive search of Figure 2, which revealed 112 domains of 20 residues each but only 5 domains of 21 residues each. The first two domains above are actually more rigid, in the sense of smaller RMS, than any of the five domains with 21 residues found exhaustively in Figure 2. The four domains found above also intersect extensively with the 5 of 21. Of the 18 residues common to the five domains with 21 residues, 10 also occur in the above four domains of 20 residues. Thus the four domains found above are close approximations to the five largest domains found exhaustively in the structure, which themselves, because of their extensive overlap, represent what is essentially one domain. As a matter of interest, the RMS fit for the entire peptide using the centroid of all 42 residues is 0.324 Å.

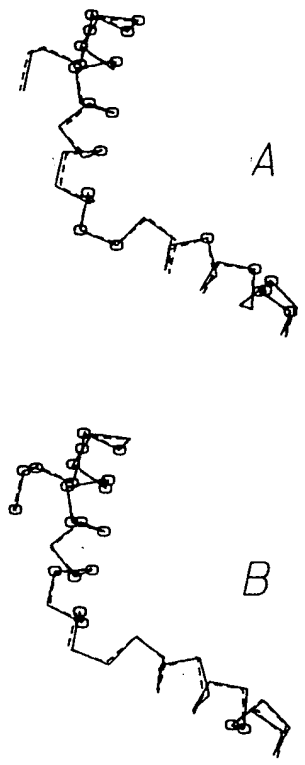

Fig. 3. The superposition of the oxy (dashed) upon the deoxy conformation of the A through C helices of the α_1 monomer of hemoglobin. **A:** Superposition using a domain found from $D_{0.30}(25)$ with an RMS fit of 0.186 Å. **B:** Superposition using a domain found from $D_{0.30}(36)$ with an RMS fit of 0.213 Å. The N-terminus is at the lower right and C-terminus of the C helix is at the upper left for both A and B views of the superimposed peptide. The graphical superpositions illustrate the differences between these two domains but show their equivalence for structural comparison.

We resume our search of the first 42 residues of the α_1 monomer of human hemoglobin to see the dependence of both $U_\varepsilon(N_s)$ and $D_\varepsilon(N_s)$ upon N_s. Repeated trials show that 36 is the smallest value for N_s that leads to a set $U_{0.30}(N_s)$ of residues common to all the largest domains of the exhaustive search of Figure 2. Because $U_{0.30}(36)$ is a domain and no residues need to be removed from it, $U_{0.30}(36)$ and $D_{0.30}(36)$ are identical.

$$U_{0.30}(36) = D_{0.30}(36) = \{20\ 27\ 29\ 30\ 33\ 41\}.$$

The complement of $D_{0.30}(36)$ is now searched for residues rigid with those already in $D_{0.30}(36)$, adding such residues one at a time to $D_{0.30}(36)$, just as in Figure 2, but now with many fewer combinations to examine. This gives the following table, in which the first column is the size of a domain and the second is the number of domains of that size for which $D_{0.30}(36)$ is a subdomain:

J	N(J)
6	1
7	25
8	245
9	1306
10	4478
11	10816
12	19290
13	26008
14	26827
15	21278
16	12953
17	5981
18	2038
19	485
20	72
21	5

No larger domains exist to 0.30 Å. The five domains of 21 residues each in the last line of the table are identical to those revealed by the exhaustive search of Figure 2.

Thus we have obtained with much less computational effort the results of the exhaustive search and have also found other slightly smaller domains which are actually more rigid. A sample of the results is shown in Figure 3, where a selected two of the domains have been used to align the first 42 residues of α_1 human hemoglobin. The peptide is aligned in (A) by superimposing the oxy and deoxy forms of the 20 residue domain with an RMS fit of 0.186 Å derived from $D_{0.30}(25)$, while the alignment in (B) results from similarly superimposing the last of the five domains listed at the bottom of Figure 2, which has an RMS fit of 0.213 Å.

The various domains that we have found can be thought of as constituting a family of closely related domains that represent one rigid object with minor variations. As seen in Figure 3, two domains of this family are practically equivalent for alignment purposes.

A Method for Finding Rigid Domains From a Subdomain of $D_\epsilon(N_s)$

Larger domains can be found from the domain $D_\epsilon(N_s)$ by including with $D_\epsilon(N_s)$ all combinations of residues within the complement of $D_\epsilon(N_s)$ that are rigid with $D_\epsilon(N_s)$. All these larger domains contain $D_\epsilon(N_s)$ as a subset. In place of $D_\epsilon(N_s)$, however, any subset of $D_\epsilon(N_s)$ can be used as the domain common to all subsequently larger domains, since any subset of $D_\epsilon(N_s)$ is also a domain. By doing so we can find domains that have specific characteristics we may wish to retain. For example, residues that are either spatially or sequentially separate from the rest of the residues in $D_\epsilon(N_s)$ can be eliminated. The subsequent search through the complement of this subdomain of $D_\epsilon(N_s)$ will then lead to larger domains within the protein that retain the desired residues of the subdomain.

We show how this works with the example of the previous section, the first 42 residues of the α_1 monomer of human hemoglobin. $D_{0.30}(25)$ includes residues 1, 2, 3, 6, 9, 13, and 18, all of which are part of the A helix or N-terminus. Removing these from $D_{0.30}(25)$ leaves a subdomain lying only within the B and C helices:

$$D_{0.30}(25)_{mod} = \{20\ 24\ 27\ 28\ 30\ 31\ 33\ 36\ 37\ 38\ 39\}.$$

A search through the complement of this subdomain of $D_{0.30}(25)$ finds three of the five largest domains found with the exhaustive search of Figure 2. These domains have more B helix (residues 20–35) and less A helix (residues 3–18) than the four domains of 20 residues each first found above. The search is as follows:

J	N(J)
11	1
12	21
13	162
14	626
15	1366
16	1780
17	1424
18	709
19	218
20	39
21	3

A modest computational effort can sample a family of domains in a polypeptide and thereby escape the exponentially increasing demands for processor time and storage space required by an exhaustive search.

RESULTS

Rigid Core of the $\alpha_1\beta_1$ Dimer of Human Hemoglobin

The non-exhaustive method for finding domains is applied here to the entire $\alpha_1\beta_1$ human hemoglobin dimer. Exhaustively finding domains in a hemoglobin dimer would be unacceptably slow. The two conformations of the dimer in which we shall look for domains are the deoxy conformation 2HHB[10] and the oxy conformation 1HHO,[11] both from the Protein Data Bank.[12] A hemoglobin dimer is con-

structed from two monomers α_1 and β_1 with 141 and 146 residues, respectively.

We first search for domains in the $\alpha_1\beta_1$ hemoglobin dimer when $\varepsilon = 0.50$ Å. After the difference matrices for both the deoxy and oxy dimers have been constructed, the difference-distance matrix is computed, and a trial value of 200 is chosen for N_s. This is shown in Figure 4, in which the 104 residues in the set $U_{0.50}(200)$ are listed at the top. The element with the largest number N_M of non-rigid residues is residue $\alpha_1$11, a lysine residue in the A helix of the α_1 monomer. This residue is removed from $U_{0.50}(200)$ and the number of non-rigid residues for each residue in the remaining subset is recalculated. The next least rigid residue is $\alpha_1$72, a histidine residue in the beginning of the α_1 E-F corner. This is then removed. Residues are successively deleted in this manner until a subset of residues remains for which each member is not rigid with at most one other residue in the subset. This subset of 86 residues has six non-rigid residue pairs: ($\alpha_1$3,$\alpha_1$8), ($\alpha_1$5,$\alpha_1$109), ($\alpha_1$9,$\beta_1$51), ($\alpha_1$10,$\beta_1$115), ($\alpha_1$103,$\alpha_1$120), and ($\beta_1$22,$\beta_1$130). Removing all the non-rigid residue pairs gives the set $D_{0.50}(200)$, which has 74 residues. The search through the complement of $D_{0.50}(200)$ gives 64 domains, each with 88 residues. These are the largest domains found with $\varepsilon = 0.50$ Å. The eight residues in these domains of 88, in addition to residues in the above-listed non-rigid pairs, are residues $\alpha_1$24, $\alpha_1$27, $\alpha_1$30, $\alpha_1$33, $\alpha_1$34, $\alpha_1$111, $\beta_1$119, and $\beta_1$124. The residues of one of the largest domains are listed at the bottom of Figure 4 along with the RMS fit of the oxy onto the deoxy domain.

We show a closely related domain in Figure 5, which was found with $\varepsilon = 0.75$ Å and $N_s = 210$. $U_{0.75}(210)$ has 196 residues and leads to a domain $D_{0.75}(210)$ with 131 residues. An exhaustive search through the complement of $D_{0.75}(210)$ finds 16 domains with 135 residues each, the four residues in addition to those of $D_{0.75}(210)$ being one from each of four non-rigid residue pairs removed from $U_{0.75}(210)$ when finding $D_{0.75}(210)$. One of these 16 domains is labeled by circles in Figure 5. The 135 residue domains include most of the center of the dimer, with 81 residues from the α_1 monomer and 54 from the β_1 monomer. We call the family of 16 domains, of which this is one representative, the rigid core of the dimer to 0.75 Å.

The deoxy residues of the chosen representative of the rigid core shown in Figure 5 superpose on the oxy residues with an RMS value of 0.333 Å. To avoid cluttering Figure 5, only the deoxy hemes have been drawn. The change of conformation of both heme pockets relative to the rigid core is very apparent in Figure 5. The domain we have called the rigid core and its 15 relatives, which differ by only several residues, form a family of intersecting domains that represent minor variations of the same structure.

Effect of Changing ε

Non-zero difference-distance matrix elements Δ_{ij} defined by Eq. (1) owe their magnitude to either actual differences γ_{ij} in tertiary structure between conformations, some of which might be from correlated differences in domain orientation, or to uncorrelated differences χ_{ij} attributable to random error in measurement. If the distance between residues i and j in conformation (1) is $D_{ij}^{(1)}$ and the distance between the same residues in conformation (2) is $D_{ij}^{(2)}$, then

$$D_{ij}^{(2)} = D_{ij}^{(1)} + \gamma_{ij} + \chi_{ij}. \tag{4}$$

Matrix elements Δ_{ij} are then

$$\Delta_{ij} = |\gamma_{ij} + \chi_{ij}|. \tag{5}$$

We will assume no correlation between the Δ_{ij} and will consider only those differences χ_{ij} from experimental error. Certainly, $\Delta_{ij} = \chi_{ij}$ if the peptide being considered is perfectly rigid. The following question arises: is the identification of a family of rigid domains critically dependent on the value chosen for ε? With the rigid core of hemoglobin as an example, the circles in Figure 6 show how the number of residues included in the domain increases with ε. The behavior of these points is the expected consequence of the limited precision of the data. When ε is chosen to be small, few residues meet the criterion, but this number increases rapidly as ε is increased. This thought can be put in quantitative terms by the following argument.

Suppose that we have a set of N residues in one conformation of a protein and we have found all the connecting distances r_{ij}. Because of limited experimental precision the positions of the residues contain small errors that are assumed to be Gaussianly distributed. We also have another conformation of the protein with the same set of N residues with different errors drawn from the same Gaussian distribution. As a result, both sets of r_{ij} contain Gaussianly distributed random errors. From Eq. (1) the probability distribution of the elements of the difference–distance matrix for this set of N residues is now

$$W(\Delta_{ij}) = \frac{2}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{\Delta_{ij}^2}{2\sigma^2}\right) \tag{6}$$

where σ is the standard deviation of the errors in the Δ_{ij}. A subset of the N residues rigid to an ε similar in value to σ will have M residues, with M less than or equal to N. Because the Δ_{ij} are assumed to be uncorrelated, the probability that any subset containing M elements of the rigid set will meet the ε criterion is

$$\left(\int_0^{\varepsilon} W(\Delta)d\Delta\right)^{M(M-1)/2} \tag{7}$$

$U_{0.50}(200)$
α_1 residues:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 19 20 31 35
36 38 39 41 42 43 60 70 71 72 102 103 104 105 106 108 109 110 113 115
116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133
β_1 residues:
15 16 17 20 22 23 32 33 34 35 37 38 50 51 52 53 55 82 107 108
109 110 111 112 113 114 115 116 117 118 120 121 122 123 125 126 127 128 129 130
131 132 133 134 135 136

J	EXCLUDED RESIDUE	N_M
104	$\alpha_1$11	24
103	$\alpha_1$72	20
102	$\alpha_1$4	19
101	$\beta_1$20	16
100	$\alpha_1$16	11
99	$\beta_1$55	8
98	$\alpha_1$12	8
97	$\alpha_1$15	7
96	$\beta_1$108	6
95	$\alpha_1$115	5
94	$\alpha_1$106	5
93	$\alpha_1$71	5
92	$\alpha_1$14	5
91	$\beta_1$38	4
90	$\alpha_1$105	4
89	$\beta_1$133	3
88	$\alpha_1$118	3
87	$\beta_1$82	2

Remaining non-rigid pairs: ($\alpha_1$3,$\alpha_1$8) ($\alpha_1$5,$\alpha_1$109) ($\alpha_1$9,$\beta_1$51) ($\alpha_1$10,$\beta_1$115) ($\alpha_1$103,$\alpha_1$120) ($\beta_1$22,$\beta_1$130)

$D_{0.50}(200)$
α_1 residues:
1 2 6 7 13 19 20 31 35 36 38 39 41 42 43 60 70 102 104 108
110 113 116 117 119 121 122 123 124 125 126 127 128 129 130 131 132 133
β_1 residues:
15 16 17 23 32 33 34 35 37 50 52 53 107 109 110 111 112 113 114 116
117 118 120 121 122 123 125 126 127 128 129 131 132 134 135 136

J	N(J)
75	20
76	184
77	1032
78	3942
79	10848
80	22180
81	34232
82	40081
83	35436
84	23292
85	11040
86	3568
87	704
88	64

An 88 residue domain with RMS = 0.240 Å
α_1 residues:
1 2 3 5 6 7 9 10 13 19 20 24 27 30 31 33 34 35 36 38
39 41 42 43 60 70 102 103 104 108 110 111 113 116 117 119 121 122 123 124
125 126 127 128 129 130 131 132 133
β_1 residues:
15 16 17 22 23 32 33 34 35 37 50 52 53 107 109 110 111 112 113 114
116 117 118 119 120 121 122 123 124 125 126 127 128 129 131 132 134 135 136

Fig. 4. The rigid core of the human hemoglobin $\alpha_1\beta_1$ dimer as found with $\varepsilon = 0.50$ Å and $N_s = 200$. One hundred and four residues have a sum S_i of at least 200. These residues form the set $U_{0.50}(200)$. Only 86 of the residues belonging to $U_{0.50}(200)$ are non-rigid with at most one other residue in $U_{0.50}(200)$. $D_{0.50}(200)$, a domain of 74 residues, is left after all 12 residues belonging to non-rigid pairs have been removed. Looking through the complement of $D_{0.50}(200)$, we find eight residues in addition to those of non-rigid pairs that are rigid with $D_{0.50}(200)$. Only one of the largest domains is shown at the bottom of the figure along with the RMS value for its oxy-deoxy superposition.

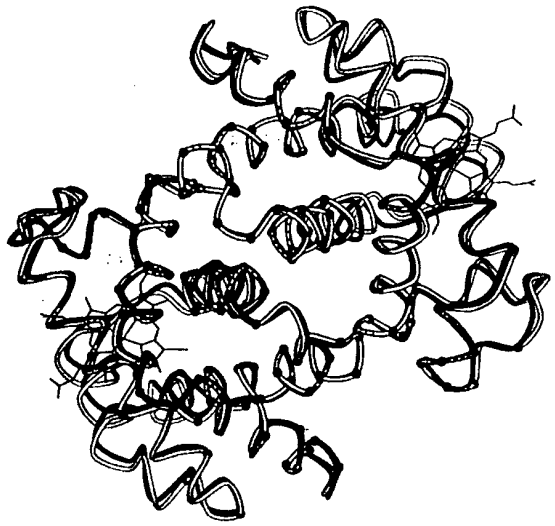

Fig. 5. Superposition of the deoxy and oxy $\alpha_1\beta_1$ dimers of human hemoglobin by aligning the residues of one of the rigid cores found with $\varepsilon = 0.75$ Å. The view is down the x-axis toward the dimer-dimer interface. The α_1 monomer is to the lower ledt, and the β_1 is to the upper right. Blue and green, the α carbon backbone of the deoxy conformation; red, the oxy conformation. About half of the same rigid core was found with visual methods by Baldwin and Chothia.[16] The RMS value for the oxy-deoxy superposition of this rigid core is 0.333 Å. Only deoxy hemes, colored brown, have been drawn in the figure. Both heme pockets clearly undergo considerable conformational change relative to the rigid core.

since there are $M(M-1)/2$ pairs of the M residues, all of which must have Δ less than ε. A total of $C(N,M)$ subsets of M residues exists in the set of N residues, so the expected number of subsets of M residues that define a domain is $C(N,M)$ times Eq. (7). We seek the largest value of M for a given ε for which we find at least one domain within the original set of N residues. Thus the largest M is the integer closest to the solution of the equation

$$C(N,M)\left(\int_0^{\varepsilon} W(\Delta)d\Delta\right)^{M(M-1)/2} = 1. \qquad (8)$$

Actually, for the hemoglobin dimer, we find that the data are fit much better if we assume that there are two disjoint subsets, one with m_1 points described by $W_1(\Delta)$ with standard deviation of σ_1 and the other with m_2 points described by $W_2(\Delta)$ with σ_2. Eq. (8) then generalizes to

$$C(N,M)\left(\int_0^{\varepsilon} W_1(\Delta)d\Delta\right)^{m_1(m_1-1)/2} \left(\int_0^{\varepsilon} W_2(\Delta)d\Delta\right)^{m_2(m_2-1)/2} = 1 \qquad (9)$$

with $M = m_1 + m_2$.

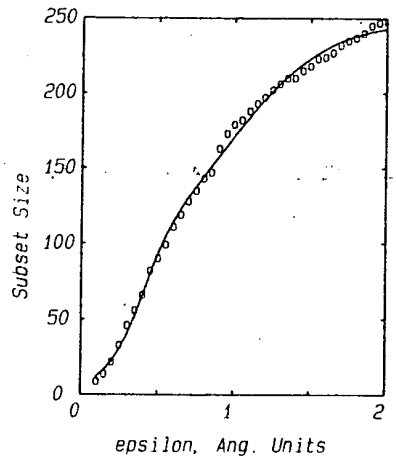

Fig. 6. The dependence of hemoglobin core size on ε. Circles mark the number of residues in a core domain for each ε as found from PDB atomic coordinates. The solid line is a best-fit of Eq. (9) with $\sigma_1 = 0.20$ Å and $\sigma_2 = 0.86$ Å to the measured points. We have assumed in this calculation that difference-distance matrix elements Δ_{ij} have a Gaussian distribution.

A best-fit curve of M versus ε obtained from the solution of Eq. (9) is shown in Figure 6. This curve was obtained by varying the parameters σ_1, σ_2, and m_1 until the sum of the squares of the deviations from the points was minimized. The values found for σ_1 and σ_2 were 0.20 Å and 0.86 Å, respectively, in the neighborhood of the experimental precision of about 0.5 Å. (The standard deviation of this curve from the points is 4.0, whereas the standard deviation found when attempting to fit with only one Gaussian subset was 25; the one-Gaussian fit was not satisfactory.) Thus even this crude theory of the dependence of the number of residues M in a domain on ε gives a reasonably good description of the observations.

In Figure 7 we show the rigid domains found with ε values of 0.25 Å (asterisks) and 0.50 Å (circles), while Figure 5 shows the 0.75 Å domain. The core structure appears to be well marked by the 0.50 Å circles. Increasing ε from 0.50 Å to 0.75 Å mainly picks up more residues in the same structure while extending the structure only slightly. Apparently the principal difference between the 0.50 and 0.75 cores is that the latter is more tolerant of errors in the data. This is in accord with Baldwin and Chothia's[16] estimate that differences between coordinates in their data were not significant unless they exceeded about 0.50 Å because of experimental uncertainty in the coordinates. This domain is definitely though sparsely marked in Figure 7 even by the 0.25 Å asterisks. Thus the identification of the gross structure of a rigid domain is not very sensitive to the value of ε for sufficiently large ε.

CONCLUSIONS

Proceeding from the premise that if rigid domains exist they should be important components of pro-

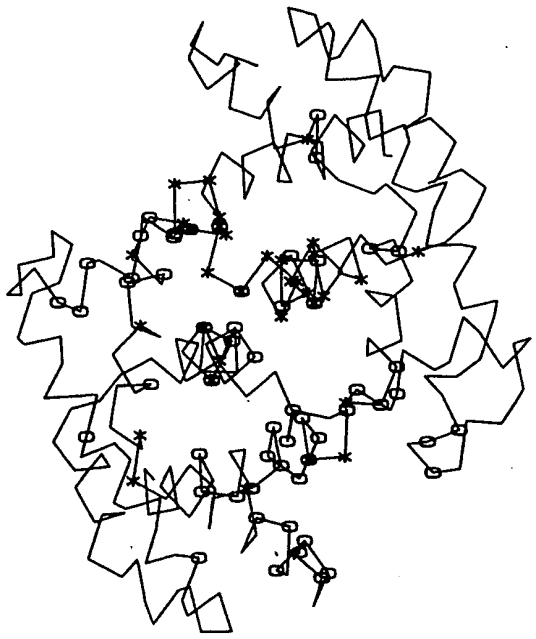

Fig. 7. A view down the x-axis of the hemoglobin dimer with residues in a rigid core for $\varepsilon = 0.25$ Å marked with asterisks and additional rigid core residues found with $\varepsilon = 0.50$ Å marked with circles. Apparently, the effect of increasing ε is to fill in secondary structures already defined by smaller values of ε.

tein structure, we have devised two methods for finding such domains, and have tried out these methods using subunits of hemoglobin. We have found that rigid domains occur in families, the members of which overlap extensively, that differ by only a few residues. The concept of a family of overlapping domains is an important generalization of the rigid-domain concept itself.

Nearly all the residues belonging to the hemoglobin dimer rigid core of Figure 5 are found within the A, B, C, G, or H helices of the α_1 and β_1 monomers. Similar structures have been noted before. Baldwin and Chothia[16] identified 68 residues that form an invariant set along the $\alpha_1\beta_1$ interface of the hemoglobin dimer. These residues are mostly the parts of the α and β B, G, and H helices and were used as a frame of reference by Baldwin and Chothia[16] from which to observe the tertiary and quaternary changes in hemoglobin. Except for residues $\beta_1 30$ and $\beta_1 31$, which are within the interior of the β_1 B helix, and residue $\beta_1 54$, a valine residue in the β_1 D helix, all are included in our family of 16 rigid core domains with $\varepsilon = 0.75$ Å. Baldwin and Chothia[16] noted as well that the α B, α C, α G, and α H helices and the β B, β D, β G, and β H helices together, except for the first few residues of the G helices and the last few residues of the H helices, remain fairly invariant between the T and R states of hemoglobin. For larger values of ε we find rigid core domains that include most of these helices but also many residues in both the α and β A helices and in the C helix of β as well. That the A, G, and H helices form a protected folding unit in apo-myoglobin has been noted by Hughson et al.[22]

The rigid core is not the only domain that can be found in the hemoglobin dimer. By removing the rigid core residues from the dimer structure and searching the remainder we can find several other smaller, independent domains associated with the heme molecules. We expect to describe these in another paper in preparation.

The primary contribution of this paper is a method to determine conserved spatial relationships. As such, it is directly applicable to analysis of complex conformational changes in proteins. Allostery is one such case; there are others, such as the calcium-triggered change in calmodulin, or the rearrangement of the hemagglutinin of influenza virus.

We have thought about the application to finding conserved cores in homologous proteins, However, that application requires substantial further development. We can calculate conserved structure given a sequence alignment, but finding the best sequence alignment for identifying conservation of structure is a another problem. The discussion of sequence alignment would take us far outside the scope of this paper.

ACKNOWLEDGMENTS

Support to the authors includes NIH 1 PO1 HL48018 to L. Ten Eyck et al., GM29458 to G. Rose, and GM11916 to B. Zimm, who also acknowledges a grant from the Wyatt Technology Corporation. This work used facilities of the San Diego Supercomputer Center under grant ASC-8902827 from the National Science Foundation. We thank Gary Ackers for helpful discussions.

REFERENCES

1. Wetlaufer, D.B. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Natl. Acad. Sci. USA 70:697–701, 1973.
2. Rossman, M.G., Liljas, A. Recognition of structural domains in globular proteins. J. Mol. Biol. 85:177–181, 1974.
3. Crippen, G.M. The tree structural organization of proteins. J. Mol. Biol. 126:315–332, 1978.
4. Rose, G.D. Hierarchic organization of domains in globular proteins. J. Mol. Biol. 134:447–470, 1979.
5. Wodak, S.J., Janin, J. Location of structural domains in proteins. Biochemistry 20:6544–6552, 1981.
6. Rashin, A.A. Location of domains in globular proteins. Nature 291:85–87, 1981.
7. Lesk, A.M., Rose, G.D. Folding units in globular proteins. Proc. Natl. Acad. Sci. USA 78:4303–4308, 1981.
8. Levitt, M., Sander, C., Stern, P.S. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. 181:423–447, 1985.
9. Zehfus, M.H., Rose, G.D. Compact units in proteins. Biochemistry 25:5759–5765, 1986.
10. Fermi, G., Perutz, M.F., Shaanan, B., Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. J. Mol. Biol. 175:159–174, 1984.
11. Shaanan, B. Structure of human oxyhaemoglobin at 2.1 Å resolution. J. Mol. Biol. 171:31–59, 1983.

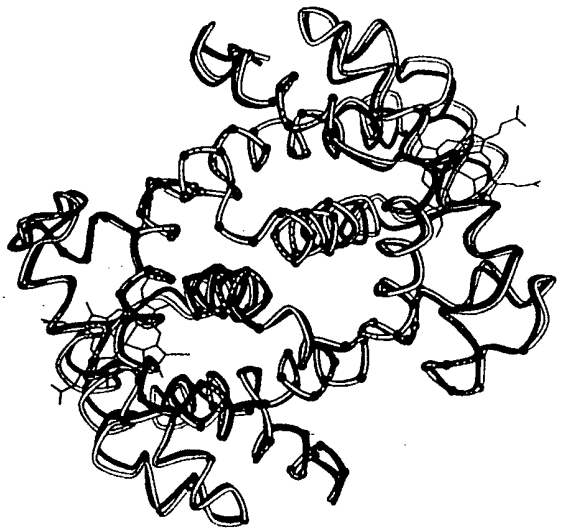

Fig. 5. Superposition of the deoxy and oxy $\alpha_1\beta_1$ dimers of human hemoglobin by aligning the residues of one of the rigid cores found with $\varepsilon = 0.75$ Å. The view is down the x-axis toward the dimer-dimer interface. The α_1 monomer is to the lower ledt, and the β_1 is to the upper right. Blue and green, the α carbon backbone of the deoxy conformation; red, the oxy conformation. About half of the same rigid core was found with visual methods by Baldwin and Chothia.[16] The RMS value for the oxy-deoxy superposition of this rigid core is 0.333 Å. Only deoxy hemes, colored brown, have been drawn in the figure. Both heme pockets clearly undergo considerable conformational change relative to the rigid core.

since there are $M(M-1)/2$ pairs of the M residues, all of which must have Δ less than ε. A total of $C(N,M)$ subsets of M residues exists in the set of N residues, so the expected number of subsets of M residues that define a domain is $C(N,M)$ times Eq. (7). We seek the largest value of M for a given ε for which we find at least one domain within the original set of N residues. Thus the largest M is the integer closest to the solution of the equation

$$C(N,M)\left(\int_0^{\varepsilon} W(\Delta)d\Delta\right)^{M(M-1)/2} = 1. \qquad (8)$$

Actually, for the hemoglobin dimer, we find that the data are fit much better if we assume that there are two disjoint subsets, one with m_1 points described by $W_1(\Delta)$ with standard deviation of σ_1 and the other with m_2 points described by $W_2(\Delta)$ with σ_2. Eq. (8) then generalizes to

$$C(N,M)\left(\int_0^{\varepsilon} W_1(\Delta)d\Delta\right)^{m_1(m_1-1)/2} \left(\int_0^{\varepsilon} W_2(\Delta)d\Delta\right)^{m_2(m_2-1)/2} = 1 \qquad (9)$$

with $M = m_1 + m_2$.

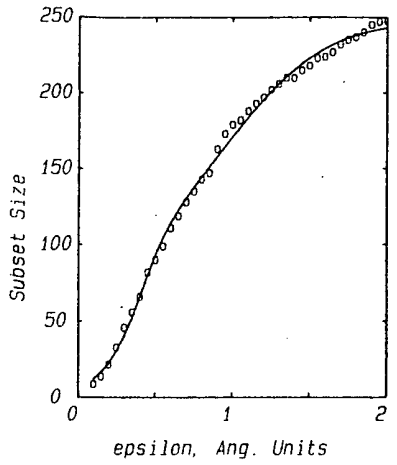

Fig. 6. The dependence of hemoglobin core size on ε. Circles mark the number of residues in a core domain for each ε as found from PDB atomic coordinates. The solid line is a best-fit of Eq. (9) with $\sigma_1 = 0.20$ Å and $\sigma_2 = 0.86$ Å to the measured points. We have assumed in this calculation that difference-distance matrix elements Δ_{ij} have a Gaussian distribution.

A best-fit curve of M versus ε obtained from the solution of Eq. (9) is shown in Figure 6. This curve was obtained by varying the parameters σ_1, σ_2, and m_1 until the sum of the squares of the deviations from the points was minimized. The values found for σ_1 and σ_2 were 0.20 Å and 0.86 Å, respectively, in the neighborhood of the experimental precision of about 0.5 Å. (The standard deviation of this curve from the points is 4.0, whereas the standard deviation found when attempting to fit with only one Gaussian subset was 25; the one-Gaussian fit was not satisfactory.) Thus even this crude theory of the dependence of the number of residues M in a domain on ε gives a reasonably good description of the observations.

In Figure 7 we show the rigid domains found with ε values of 0.25 Å (asterisks) and 0.50 Å (circles), while Figure 5 shows the 0.75 Å domain. The core structure appears to be well marked by the 0.50 Å circles. Increasing ε from 0.50 Å to 0.75 Å mainly picks up more residues in the same structure while extending the structure only slightly. Apparently the principal difference between the 0.50 and 0.75 cores is that the latter is more tolerant of errors in the data. This is in accord with Baldwin and Chothia's[16] estimate that differences between coordinates in their data were not significant unless they exceeded about 0.50 Å because of experimental uncertainty in the coordinates. This domain is definitely though sparsely marked in Figure 7 even by the 0.25 Å asterisks. Thus the identification of the gross structure of a rigid domain is not very sensitive to the value of ε for sufficiently large ε.

CONCLUSIONS

Proceeding from the premise that if rigid domains exist they should be important components of pro-

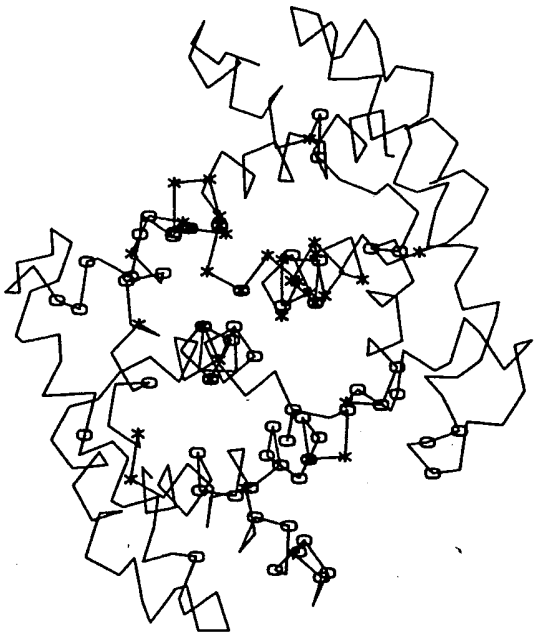

Fig. 7. A view down the x-axis of the hemoglobin dimer with residues in a rigid core for $\varepsilon = 0.25$ Å marked with asterisks and additional rigid core residues found with $\varepsilon = 0.50$ Å marked with circles. Apparently, the effect of increasing ε is to fill in secondary structures already defined by smaller values of ε.

tein structure, we have devised two methods for finding such domains, and have tried out these methods using subunits of hemoglobin. We have found that rigid domains occur in families, the members of which overlap extensively, that differ by only a few residues. The concept of a family of overlapping domains is an important generalization of the rigid-domain concept itself.

Nearly all the residues belonging to the hemoglobin dimer rigid core of Figure 5 are found within the A, B, C, G, or H helices of the α_1 and β_1 monomers. Similar structures have been noted before. Baldwin and Chothia[16] identified 68 residues that form an invariant set along the $\alpha_1\beta_1$ interface of the hemoglobin dimer. These residues are mostly the parts of the α and β B, G, and H helices and were used as a frame of reference by Baldwin and Chothia[16] from which to observe the tertiary and quaternary changes in hemoglobin. Except for residues $\beta_1 30$ and $\beta_1 31$, which are within the interior of the β_1 B helix, and residue $\beta_1 54$, a valine residue in the β_1 D helix, all are included in our family of 16 rigid core domains with $\varepsilon = 0.75$ Å. Baldwin and Chothia[16] noted as well that the α B, α C, α G, and α H helices and the β B, β D, β G, and β H helices together, except for the first few residues of the G helices and the last few residues of the H helices, remain fairly invariant between the T and R states of hemoglobin. For larger values of ε we find rigid core domains that include most of these helices but also many residues in both the α and β A helices and in the C helix of β as well. That the A, G, and H helices form a protected folding unit in apo-myoglobin has been noted by Hughson et al.[22]

The rigid core is not the only domain that can be found in the hemoglobin dimer. By removing the rigid core residues from the dimer structure and searching the remainder we can find several other smaller, independent domains associated with the heme molecules. We expect to describe these in another paper in preparation.

The primary contribution of this paper is a method to determine conserved spatial relationships. As such, it is directly applicable to analysis of complex conformational changes in proteins. Allostery is one such case; there are others, such as the calcium-triggered change in calmodulin, or the rearrangement of the hemagglutinin of influenza virus.

We have thought about the application to finding conserved cores in homologous proteins, However, that application requires substantial further development. We can calculate conserved structure given a sequence alignment, but finding the best sequence alignment for identifying conservation of structure is a another problem. The discussion of sequence alignment would take us far outside the scope of this paper.

ACKNOWLEDGMENTS

Support to the authors includes NIH 1 PO1 HL48018 to L. Ten Eyck et al., GM29458 to G. Rose, and GM11916 to B. Zimm, who also acknowledges a grant from the Wyatt Technology Corporation. This work used facilities of the San Diego Supercomputer Center under grant ASC-8902827 from the National Science Foundation. We thank Gary Ackers for helpful discussions.

REFERENCES

1. Wetlaufer, D.B. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Natl. Acad. Sci. USA 70:697–701, 1973.
2. Rossman, M.G., Liljas, A. Recognition of structural domains in globular proteins. J. Mol. Biol. 85:177–181, 1974.
3. Crippen, G.M. The tree structural organization of proteins. J. Mol. Biol. 126:315–332, 1978.
4. Rose, G.D. Hierarchic organization of domains in globular proteins. J. Mol. Biol. 134:447–470, 1979.
5. Wodak, S.J., Janin, J. Location of structural domains in proteins. Biochemistry 20:6544–6552, 1981.
6. Rashin, A.A. Location of domains in globular proteins. Nature 291:85–87, 1981.
7. Lesk, A.M., Rose, G.D. Folding units in globular proteins. Proc. Natl. Acad. Sci. USA 78:4303–4308, 1981.
8. Levitt, M., Sander, C., Stern, P.S. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. 181:423–447, 1985.
9. Zehfus, M.H., Rose, G.D. Compact units in proteins. Biochemistry 25:5759–5765, 1986.
10. Fermi, G., Perutz, M.F., Shaanan, B., Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. J. Mol. Biol. 175:159–174, 1984.
11. Shaanan, B. Structure of human oxyhaemoglobin at 2.1 Å resolution. J. Mol. Biol. 171:31–59, 1983.

12. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol: 112:535–542, 1977.
13. Vriend, G., Sander, C. Detection of common three-dimensional substructures in proteins. Proteins 11:52–58, 1991.
14. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233:123–138, 1993.
15. Dickerson, R.E., Geis, I. "Hemoglobin: Structure, Function, Evolution, and Pathology." Menlo Park, CA: Benjamin/Cummings Publishing Company, 1983.
16. Baldwin, J., Chothia, C. Haemoglobin: The structural changes related to ligand binding and its allosteric mechanism. J. Mol. Biol. 129:175–220, 1979.
17. Frauenfelder, H., Hartmann, H., Karplus, M., Kuntz, I.D., Jr., Kuriyan, J., Parak, F., Petsko, G.A., Ringe, D., Tilton, R.F., Jr., Connolly, M.L., Max, N. Thermal expansion of a protein. Biochemistry 26:254–261, 1987.
18. Richards, F.M., Kundrot, C.E. Identification of structural motifs from protein coordinate data: Secondary structure and first-level super secondary structure. Proteins 3:71–84, 1988.
19. Bystroff, C., Kraut, J. Crystal structure of unliganded *Escherichia coli* dihydrofolate reductase. Ligand-induced conformational changes and cooperativity in binding. Biochemistry 30:2227–2239, 1991.
20. Kabsch, W. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A32:922–923, 1976.
21. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A34: 827–828, 1978.
22. Hughson, F.M., Wright, P.E., Baldwin, R.L. Structural characterization of a partly folded apomyoglobin intermediate. Science 249:1544–1548, 1990.

PROTEINS: Structure, Function, and Genetics 11:52–58 (1991)

Detection of Common Three-Dimensional Substructures in Proteins

Gerrit Vriend and Chris Sander
European Molecular Biology Laboratory, D-6900 Heidelberg, Federal Republic of Germany

ABSTRACT **We present a fully automatic algorithm for three-dimensional alignment of protein structures and for the detection of common substructures and structural repeats. Given two proteins, the algorithm first identifies all pairs of structurally similar fragments and subsequently clusters into larger units pairs of fragments that are compatible in three dimensions. The detection of similar substructures is independent of insertion/deletion penalties and can be chosen to be independent of the topology of loop connections and to allow for reversal of chain direction. Using distance geometry filters and other approximations, the algorithm, implemented in the WHAT IF program, is so fast that structural comparison of a single protein with the entire database of known protein structures can be performed routinely on a workstation. The method reproduces known non-trivial superpositions such as plastocyanin on azurin. In addition, we report surprising structural similarity between ubiquitin and a (2Fe-2S) ferredoxin.**

Key words: protein structure comparison, superposition, clustering, folding units, sequence alignment

INTRODUCTION

Comparison of protein structures has many areas of application. Three-dimensional similarity can be used to produce protein alignments in cases where sequence similarity is so weak that sequence alignment programs fail.[1] Structure-based sequence alignment can reveal evolutionary relationships and provide the basis for the construction of phylogenetic trees.[2] Multiple alignment of structures naturally leads to the definition of a common structural core of a protein family,[3] to the identification of structurally important conserved contact regions,[4] and to the detailed study of residue replacements in conserved structural context.[5]

The principal difficulty in comparing three-dimensional protein structures is that of identifying structurally equivalent residues. Once a list of equivalent residues is known, elegant solutions to the problem of optimal superposition in 3-D[6] can be used to produce explicit coordinates of one protein in the framework of the other. Superficially, the equivalencing problem is similar to the problem of one-dimensional alignment of amino acid sequences. There is, however, an added complication in that clusters of residues locally similar in three-dimensional space may involve chain regions separated by many residues, i.e., arranged non-locally in sequence space. It is therefore not sufficient to compare one-dimensional neighborhoods in sequence space, but also necessary to compare three-dimensional neighborhoods in real space. For this reason, one-pass dynamic programming algorithms are not suitable for this problem.

Several authors have invented generalizations of sequence alignment algorithms in order to solve the 3-D equivalencing problem. For example, Taylor and Orengo[7] first define a local measure of similarity between any two sequence positions in two proteins by aligning the contact environments of each residue in protein A with that of each residue in protein B, using a dynamic programming algorithm. Subsequently, they solve the one-dimensional alignment problem in terms of new local similarities derived from the first step, again by dynamic programming. The algorithm can be thought of as solving the problem of aligning two contact maps (or distance plots), allowing insertions and deletions but adhering strictly to the sequential order of residues along the chain. This method is conceptionally neat and works well, but it is costly in computer time, as the algorithm is of order $N(A)^2N(B)^2$, where N(X) is the chain length of protein X. Sali and Blundell[8] use a Monte Carlo method, simulated annealing, to deal with the complexity of optimizing structural superposition, whereas Zuker[9] uses a dynamic programming algorithm.

Several other known methods for protein structure comparison are not based on generalizations of sequence alignment algorithms, but use a variety of iterative schemes to optimize superposition.[10–19] These methods have been extensively used for (closely) related structures. However, they each suffer from one or more of the following drawbacks: (1)

Received August 27, 1990; revision accepted February 1, 1991.

Address reprint requests to either Chris Sander or Gerrit Vriend, European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 102209, D-6900 Heidelberg, Germany

large insertions and deletions are difficult to recognize; (2) only sequential alignments can be detected; (3) neither the occurrence of multiple copies of a motif nor spatial similarity in spite of different loop connections are detectable; (4) manual initial alignment is required; and (5) massive CPU resources are needed.

Here, an algorithm for protein structure comparison is proposed that overcomes these problems. The procedure consists of three steps. One, selection of sequence-local fragments that superpose well to create a diagonal plot. This is done efficiently using distance geometry criteria[20] followed by a fast algorithm for 3-D superposition.[6] Two, cluster analysis on pairs of fragments in order to identify larger structural units. Three, final optimization of the set of equivalent residues by minor trimming and extension and final optimal superposition of coordinates.

METHODS

From Sequence Alignment to 3-D Alignment

The point of departure of our algorithm is a so-called diagonal plot, a standard device for the graphical representation of sequence alignments. The two axes of such a rectangular plot are the sequences of the two proteins. A diagonal line segment, i.e., a trace inclined at 45° represents the similarity of a stretch of a certain length in protein A with a stretch of the same length in protein B without insertions or deletions, e.g., the similarity between two beta strands. The classical sequence alignment problem is that of finding an overall optimal path through the diagonal plot, connecting diagonal line segments, such that the overall similarity, i.e., residue similarity summed over all residues in all fragment pairs, is optimal, with sequential order strictly maintained. Sequential order is satisfied for two pairs of matching fragments, say, A1/B1 and A2/B2, if the fragments occur in the same order in both proteins, i.e., either A1<A2 and B1<B2 or A1>A2 and B1>B2, but not in mixed order like A1<A2 and B1>B2 or A1>A2 and B1<B2, where < means "comes before" in sequence and > means "comes after." See Figure 1A for an example.

In 3-D alignment, an optimal sum over fragment pair similarities alone does not guarantee that the matched segment pairs are part of similar substructures. For example, suppose that fragments A1 and A2 each are similar in shape to fragments B1 and B2, respectively. If, however, the spatial relationship of A1 and A2 in protein A differs from that of B1 and B2 in protein B, then the fused substructure A1 + A2 is not similar to B1 + B2; see Figure 1B for an example. The generalization of this argument from two to N fragment pairs leads to a clustering algorithm in which a new fragment pair Ai/Bi is added to an existing cluster of pairs if the spatial relationship between Ai and the A fragments in the cluster is similar to that of Bi and the B fragments already in the cluster. The requirement of strict sequential order of equivalenced fragments can be dropped. For example, one can allow A2/B2 to join in a cluster with A1/B1 even if A1 < A2 and B2 > B1 in sequence.

Technically, similarity of spatial relationship can be evaluated either in terms of explicit 3-D superposition or in terms of intrafragment distances. For example, one could simply determine the optimal superposition of A1 + A2 as one piece onto B1 + B2 and apply a cutoff in positional rms (root mean square) deviation as a criterion for joining these pairs into a common cluster. Alternatively, on could compare a set of alpha-carbon distances within A1 + A2 with an equivalent set within B1 + B2. A much more efficient test of spatial relationship can be made in terms of quantities already calculated in the production of the diagonal plot. This is our key technical point. The idea is to assess the similarity of spatial orientation between a pair A1/B1 and another pair A2/B2 by comparing the rotation operator attached to each pair comparison. The union of fragments A1 + A2 in protein A is similar to the union of fragments B1 + B2 if and only if the rotational transformation that best superposes A1 onto B1 is similar to the one that best superposes A2 onto B2. The comparison of operators can be performed by multiplying one operator by the inverse of the other and quantifying the departure of the result from the unit operator. Technical details of our method are given in the next three subsections.

Diagonal Plot

The first step in the creation of the diagonal plot is the comparison of fragments in the two proteins to be aligned. All fragments with a certain minimum length from one protein are compared with all fragments of the same length from another protein. Fragment pairs of similar structure are retained and reported in the diagonal plot as diagonal traces. In order to save computer time, a geometrical filter is applied to each pair of fragments in terms of intrafragment distances. If two fragments have the same structure, they will also have the same set of internal distances. So, if distance criteria are violated, the two fragments cannot have the same structure and need not be compared in more detail. However, the converse is not true. Even if distance sets are similar, the structures may be significantly different. Therefore the (fast) comparison of sets of distances has to be followed by a (slower) explicit three-dimensional superposition in order to eliminate false positives.

The comparison of fragment geometry in terms of internal alpha-carbon distances is done by a method similar to that of Jones and Thirup,[20] except that only the distances from the first alpha-carbon atom to the last five alpha-carbon atoms in the same frag-

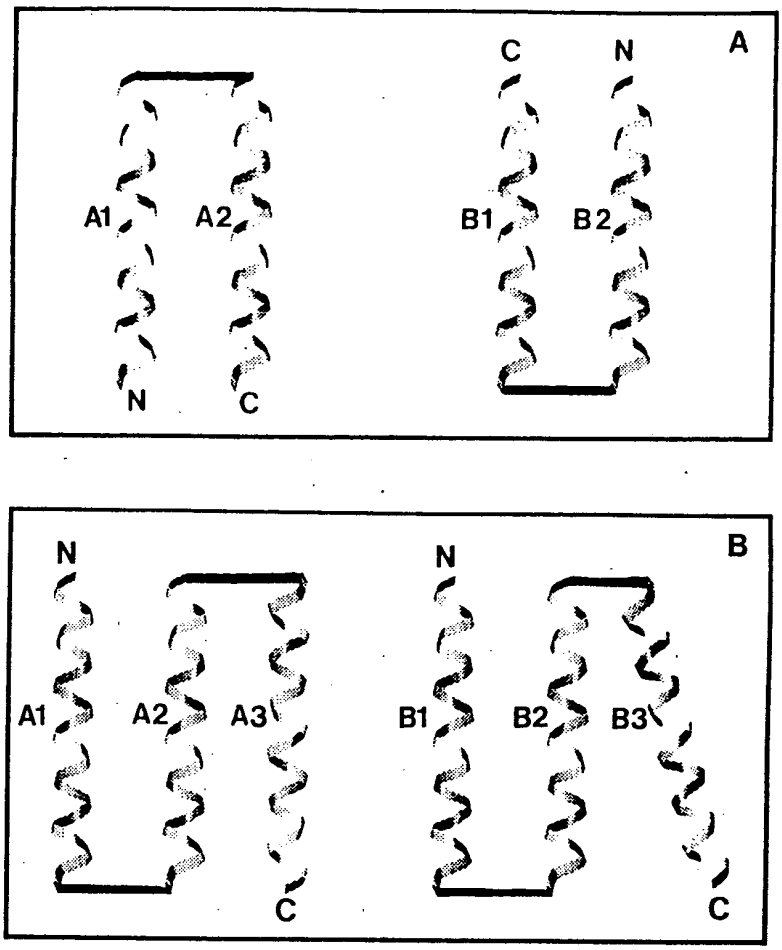

Fig. 1 Schematic example of comparing two alpha-helical protein structures, A and B, in which helix A1 is similar to helix B1, A2 to B2 and A3 to B3. **A.** The cores of the two structures superpose well, even though the interhelical connections are different. **B.** Helices A1 + A2 superpose well on B1 + B2. The third fragment pair, A3/B3, cannot join the cluster made by the other two helix pairs because the third helix has a different orientation in the two proteins.

ment are used. Using more than five distances did not appreciably improve the selectiveness of the filter. A pair of fragments is rejected if two equivalent alpha-carbon distances differ by more than a specified cutoff. To be sure that no pair will be rejected spuriously, this cutoff should be set at two times the maximum acceptable coordinate error after 3-D superposition of the fragments. The length of the shortest fragments compared is normally set at 10 to 15 residues. Using shorter fragments gives rise to an excessive number of matched fragments; using a longer minimum length tends to reduce the number of hits unacceptably.

In a second step, pairs of fragments that are not rejected by the distance geometry filter are superposed using the least-square algorithm of Kabsch.[6] This is straightforward as the two fragments in a pair have the same length. For each fragment pair, the goodness of fit is evaluated in terms of the root mean square distances, Drms, and the largest distance, Dmax, between equivalent alpha-carbon atoms, after optimal superposition. Two fragments are considered to be sufficiently similar if these distance values are below specified upper limits, typically 2.0 Å for Drms and 3.8 Å for Dmax (tighter limits should be used for very similar proteins).

In a third step, accepted fragment pairs are elongated: one residue is added at the C-terminus of both fragments and the longer fragments are again superposed. This process is repeated until the next addition would lead to violation of the upper limits. Additional computer time is saved by avoiding the comparison of fragments that are entirely helical, as every helix always fits every other equally long or longer helix: fragments are only compared if they contain at least four non-helical residues. Secondary structure assignments are taken from the DSSP dictionary.[21] Also, fragment pairs are skipped that would be subsets of already stored fragments. Together, these three empirically developed steps produce very useful diagonal plots as input to the cluster analysis, with great economy of computer time.

Cluster Analysis

In order to assemble pairs of fragments into a pair of larger units, we use a simple incremental clustering procedure. In each step, one needs to determine if a pair of fragments can join a cluster. The most plausible way would be to simply add the fragment pair to the cluster and to perform a complete 3-D superposition on the entire set of equivalenced residues to see if the positional errors are less than some preset limit. This would, however, be a rather slow process, so a filter that determines if a pair of fragments could potentially join the cluster is needed. The filter we use assesses whether the fragment pairs A1/B1 and A2/B2 can be joined into a larger pair, i.e., if A1 + A2, taken as a rigid body, can be superposed well onto B1 + B2 (Fig. 1). This is done in two steps making use of quantities already calculated during the generation of the diagonal plot.

First, we check that the distance between the centers of mass of A1 and A2 is similar to the distance between the centers of mass of B1 and B2. If these intraprotein distances are too dissimilar, then there cannot be a good superposition of A1 + A2 onto B1 + B2. Second, the rotation matrix of the optimal superposition A1/B1 is compared with that of A2/B2. This is done by multiplying one superposition rotation matrix (R1) with the inverse of the other (R2) and quantifying the departure from the unit matrix in terms of the resulting net rotation angle δ given by

$$\cos \delta = \frac{1}{2}[\mathit{trace}(R_1 \cdot R_2^{-1}) - 1].$$

The discrepancy angle δ is equal to zero if R1 and R2 represent identical rotations. If δ is above a cutoff value, typically 0.2–0.3 radians, the two rotations are considered dissimilar and the fragments cannot be merged. The rationale behind this is as follows: if two proteins are perfectly superposable, then every pair of equivalent substructures is also perfectly superposable, with the same rotational component of the superposition transformation; deviations from perfect superposability can be measured in terms of deviations in the rotational component. Because the same reasoning does not hold for the translational component of the transformation, we use the vector between the centers of mass, as described above.

In order to determine the largest cluster(s) for a given protein pair, each pair of fragments should be used in turn to start a new clustering process. In general, this implies N^2 comparisons of pairs of fragment pairs, given N pairs of fragments in the diagonal plot. In practice, it is often satisfactory to terminate the search as soon as a sufficiently large cluster is found, say, exceeding the size of a minimal folding unit (> 40–50 residues) or, say, containing half of all residues in one of the proteins. If the two proteins have a measurable degree of similarity, it is likely that the fragment pairs near the main diagonal will provide the largest cluster. Therefore, in practice we search for clusters along the main diagonal first and terminate on cluster size, reducing the complexity of the clustering procedure from order N^2 to order N.

Final Adjustment and Equivalencing

Creation of the diagonal plot and clustering of pairs of fragment in principle solves the problem of structural 3-D alignment. However, for practical reasons having to do with some of the time-saving approximations, a final pruning and fine-tuning of the largest clusters is performed. These reasons are: (1) pair comparison of fragments of length, say, less than 10 residues was avoided in the initial step, but in the final superposition, such short fragments could be interesting, provided they fit into the overall context; (2) in the clustering procedure, a new fragment pair was only compared with the starting member of the cluster, so one is not yet sure that the distance criteria are fulfilled for the entire cluster; and (3) it may be of interest to detect additional segments that are part of the cluster only if their chain direction is reversed.

The largest clusters of pairs of fragments are selected and fine-tuned with an iterative procedure similar in part to that of Rao and Rossmann.[17] First, the fragments are optimally superposed, such that Drms, the average positional deviation, is minimal. Subsequently, the list of equivalenced residues is re-examined and adjusted according to the criteria discussed below and a new overall transformation and Drms are determined. The process is iterated until no further adjustment is required. This termination condition is normally fulfilled within six to nine cycles. In the equivalencing pass of the final optimization a pair of residues is accepted: (1) if all equivalenced alpha-carbon positions are within Dmax of each other and; (2) if the pair of residues is part of two consecutive stretches of minimal length (say, 5 residues), acceptable according to (1). Optionally, fragments are allowed to run sequentially in opposite directions. The final cluster is reported after optimal superposition as a list of equivalenced residues, i.e., as the structure-derived sequence alignment.

RESULTS

As a test of the method, several well-known comparisons were redone: two hemoglobin chains, plastocyanin-azurin, and the two domains of rhodanese. In addition, we report discovery of an unexpected structural similarity: ubiquitin-ferredoxin. For the alpha and beta chain of hemoglobin[22] (Fig. 2), our alignment agrees with that of Lesk and Chothia.[3] For plastocyanin[23] azurin[24] (Fig. 3), our result agrees with the alignment by Adman.[25] A known

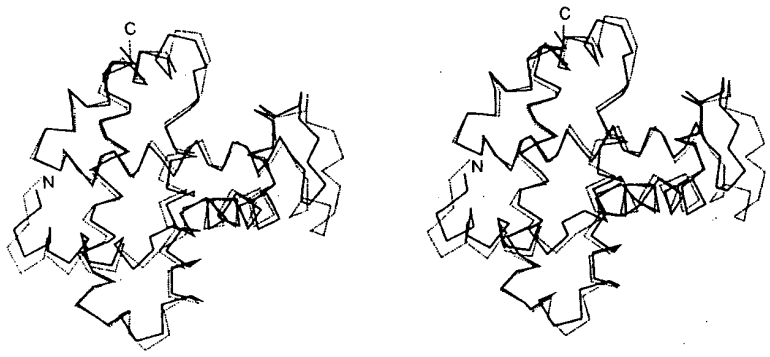

Fig. 2. Human deoxyhemoglobin beta chain (dashed lines) superposed on the alpha chain (solid lines). Stereo view. N terminus and C terminus labelled.

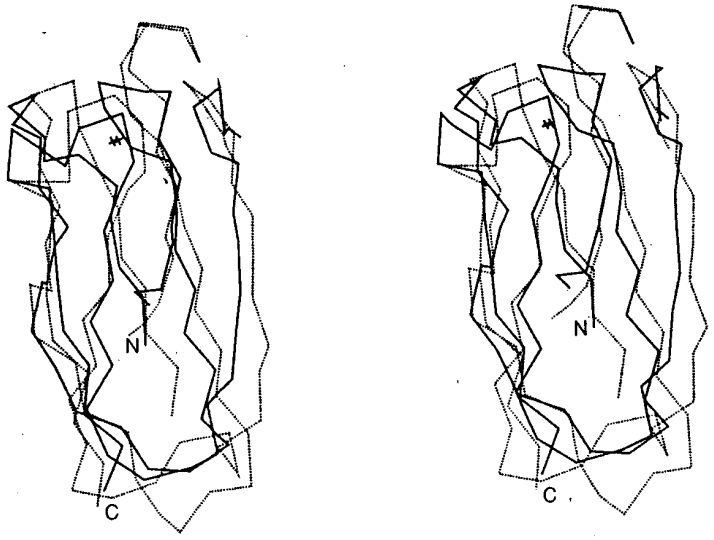

Fig. 3. Plastocyanin (dashed lines) superposed on azurin (solid lines). Two non-homologous loops (residues 43–62 in plastocyanin; residues 51–86 in azurin) are not shown, for clarity. The two bound copper ions are indicated by crosses. Stereo view.

example of internal duplication is bovine liver rhodanese.[1] This molecule is composed of two structurally similar domains, with no detectable sequence similarity between them. The cores of these two domains are almost identical, but the loop regions vary in length. The fragment match diagonal plot (Fig. 4), comparing the first with the second domain, has traces near the main diagonal that can be merged into one large cluster, corresponding to the superposition of the two domains (Fig. 5). The derived alignment is essentially identical to the one determined by the Ploegman et al.[1] with at most one residue more or less equivalenced at the ends of fragments.

A first database scan turned up several new structural similarities. One example is the pair ubiquitin[26] / ferredoxin[27] (Fig. 6) Ubiquitin, a 76 residue protein, is involved in protein breakdown via covalent conjugates, whereas ferredoxin, with 98 residues, functions as an electron carrier in the pho-

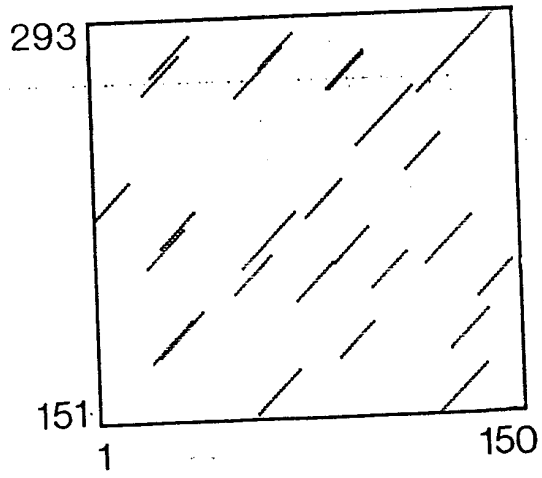

Fig. 4. Diagonal plot of fragment similarities in rhodanese, between the first domain (residues 1–150) and the second domain (residues 151–293), as used in the first step of the algorithm. Each trace corresponds to a fragment pair, which may or may not fit with the overall domain comparison.

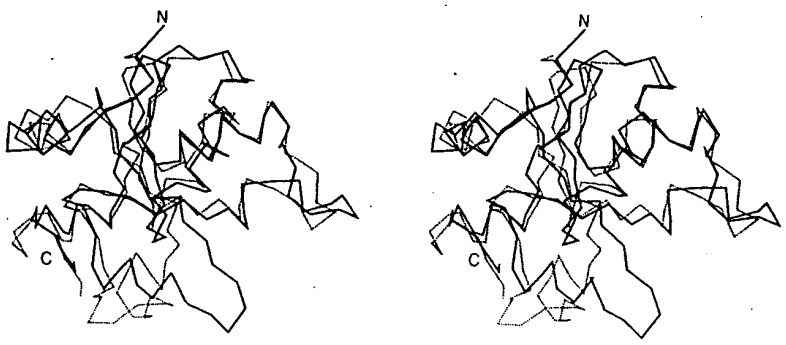

Fig. 5. Rhodanese C-terminal domain (dashed lines) superposed on its N-terminal domain (solid lines). The terminal 4 (7) residues are not shown, for clarity. Stereo view.

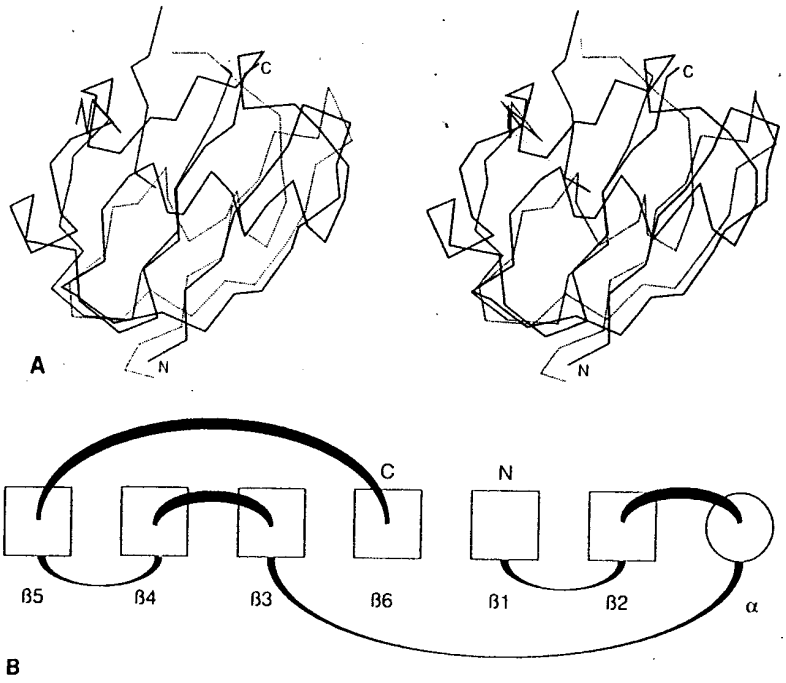

Fig. 6. **A.** Ferredoxin (dashed, 1UBQ) superposed on ubiquitin (solid lines, 3FXC). Stereo view. Two loops in ferredoxin for which no equivalent loops are present in ubiquitin are removed (top left), and replaced by thin dashed lines, for clarity. The superposition is generated by superposing the following fragments, equivalenced by the algorithm (1ubq range / 3fxc range): M1-L8/Y3-E10; T21-A28/N15-E31; G47-S57/G74-T84; E64-V70/D86-H92. **B.** Topology scheme for ferredoxin and ubiquitin. Circle: alpha helix; squares: extended strands. The alpha helix lies across the open hand formed by the beta strands 1,2,3,4, and 6. Strands 2, 1, 6, 3 and 4 form a sheet in which the irregular strand 5 does not participate. The two domains have exactly the same topology.

toreduction of cytochrome *c*. Surprisingly, the three-dimensional structures are remarkably similar. The overall rms deviation of 47 out of maximally 76 equivalenced alpha carbon atoms is 2.1 Å. Both structures can be described as a hand of five beta strands holding a short beta strand and an alpha helix in the center. There is no obvious analogy of protein function and, apparently, the structural similarity had gone undetected. Perhaps ubiquitin and ferredoxin do have a common ancestor. Alternatively, the ferredoxin and ubiquitin "beta-grasp" domain may be an energetically favored folding unit.

CONCLUSION

Our algorithm provides a novel tool for the comparison of protein structures with the options of allowing for altered loop topology and for reversal of chain direction. The entire procedure is fully automatic and can be used in a routine manner. The method is so fast that the comparison of one single structure with all known structures is possible with only a few hours CPU usage on a workstation. Large insertions or deletions or many insertions or deletions are no problem. The method can be used in any context where structural alignment is useful, e.g., to

determine reliable (structure based) sequence alignments, to aid in the definition of structural cores of protein families, and to find common three-dimensional folding units.

The method is implemented as an option in the molecular modeling and drug design program WHAT IF,[28] facilitating immediate visualization by computer graphics. WHAT IF is written in FORTRAN 77, with graphics drivers for Evans and Sutherland and Silicon Graphics computers. The program is available from G.V. for a minimal fee. Send electronic mail to VRIEND@EMBL-Heidelberg.DE on internet for information.

ACKNOWLEDGMENTS

We thank Georg Tuparev, Anna Tramontano, and Ruben Abagyan for very helpful discussions; colleagues in the EMBL Biocomputing groups for providing useful test cases; W.L. Kabsch for use of the program U3B; Evans and Sutherland and Silicon Graphics for technical support; and many crystallographers for depositing their coordinates in the Protein Data Bank.[29]

REFERENCES

1. Ploegman, J.H., Drenth, J., Kalk, K. H., Hol, W.G.J. Structure of bovine liver rhodanese. J. Mol. Biol. 123:557–594, 1978.
2. Johnson, M.S., Sali, A., Blundell, T.L. Phylogenetic relationships from three-dimensional protein structures. Meth. Enzymol. 183:670–690, 1990.
3. Lesk, A.M., Chothia, C. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. J. Mol. Biol. 136:225–270, 1980.
4. Godzik, A., Sander, C. Conservation of residue interactions in a family of Ca-binding proteins. Prot Eng. 2:589–596, 1989.
5. Bordo, D., Argos, P. Evolution of protein cores. Constraints in point mutations as observed in globin tertiary structures. J. Mol. Biol. 211:975–988, 1990.
6. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst. A 34:8274–828, 1978.
7. Taylor, W.R., Orengo, C.A. Protein structure alignment. J. Mol. Biol. 208:1–22, 1978.
8. Sali, A., Blundell, T.L. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J. Mol. Biol. 212:403–428, 1990.
9. Zuker, M., Samorjai, R.L., The alignment of protein structures in three dimensions. Bull. Math. Biol. 51:55–78, 1989.
10. Levine, M., Stuart, D., Williams, J. A method for the systematic comparison of the three-dimensional structures of proteins and some results. Acta Cryst. A40:600–610, 1984.
11. Remington, S.J., Matthews, B.W. A systematic approach to the comparison of protein structures. J. Mol. Biol. 140: 77–99, 1980.
12. Remington, S.J., Matthews, B.W. A general method to assess similarity of protein structures, with applications to T4 bacteriophage lysozyme. Proc. Natl. Acad. Sci 175: 2180–2184, 1978.
13. Lesk, A.M. Detection of 3-D patterns of atoms in chemical structures. Comm. ACM 22:219–224, 1979.
14. Brint, A.T., Davies, H.M., Mitchell, E.M., Willet, P. Rapid geometric searches in protein structures. J. Mol. Graph. 7:48–53, 1989.
15. Barnton, G.J., Sternberg, M.J.E. LOPAL and SCAMP: techniques for the comparison and display of protein structures. J. Mol. Graph. 6:190–196, 1988.
16. Rossmann, M.G., Argos, P. Exploring structural homology of proteins. J. Mol. Biol. 105:75–95, 1976.
17. Rao, S.T., Rossmann, M.G. Comparison of super-secondary structures in proteins. J. Mol. Biol. 76:241–256, 1973.
18. Abagyan, R.A., Maiorov, V.N. A simple qualitative representation of polypeptide chain folds: Comparison of protein tertiary structures. J. Biomol. Struc. Dyn. 5:1267–1279, 1988.
19. Abagyan, R.A., Maiorov, V.N. An automatic search for similar spatial arrangements of alpha helices and beta strands in globular proteins. J. Biomol. Struc. Dyn. 6: 1045–1060, 1989.
20. Jones, T.A., Thirup, S. Using known fragments in protein model building and crystallography. EMBO J. 5:819–822, 1986.
21. Kabsch, W., Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22:2577–2637, 1983.
22. Fermi, G., Perutz, M.F., Shaanan, B., Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. J. Mol. Biol. 175:159–174, 1984.
23. Guss, J.M., Freeman, H.C., Structure of oxidized popular plastocyanin at 1.6 Å resolution. J. Mol. Biol. 169:521–563, 1983.
24. Baker, E.N., Structure of azurin from alcaligenes denitrificans. Refinement at 1.8 Å resolution and comparison of the two crystallographically independent molecules. J. Mol. Biol. 203:1071–1095, 1988.
25. Adman, E.T. Metalloproteins. P.M. Harrison, ed. Verlag Chemie Weinheim, 1985, Part 1, chapter 1, pp. 1–42.
26. Vijay-Kumar, S., Bugg, C.E., Cook, W.J. Structure of ubiquitin refined at 1.8 Å resolution, J. Mol. Biol. 194:531–544, 198.
27. Tsukihara, T., Fukuyama, K. Nakamura, M., Katsube, Y., Tanaka, N., Kakudo, M., Wada, K., Hase, T., Matsubara, T., Structure of a (2Fe-2S) ferredoxin from spirulina platensis. Main chain fold and location of side chains at 2.5 Å resolution. J. Biochem. (Tokyo) 90:1763–1773, 1981.
28. Vriend, G. WHAT IF: A Molecular Modeling and Drug Design Program. J. Mol. Graphics 8:52–56, 1990.
29. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. J. Mol. Biol. 112:535–542, 1977.

Protein Science (1996), 5:262–269. Cambridge University Press. Printed in the USA.

Yeast heat shock transcription factor N-terminal activation domains are unstructured as probed by heteronuclear NMR spectroscopy

HO S. CHO,*[1] COREY W. LIU,*[1] FRED F. DAMBERGER,[2] JEFFREY G. PELTON,[3] HILLARY C. M. NELSON,[3,4] AND DAVID E. WEMMER[1,3]

[1] Department of Chemistry, University of California, Berkeley, California 94720
[2] Biophysics Graduate Group, University of California, Berkeley, California 94720
[3] Structural Biology Division, Lawrence Berkeley Laboratory, 1 Cyclotron Road, Berkeley, California 94720
[4] Department of Molecular and Cell Biology, University of California, Berkeley, California 94720

(RECEIVED September 19, 1995; ACCEPTED November 21, 1995)

Abstract

The structure and dynamics of the N-terminal activation domains of the yeast heat shock transcription factors of *Kluyveromyces lactis* and *Saccharomyces cerevisiae* were probed by heteronuclear $^{15}N\{^1H\}$ correlation and $^{15}N\{^1H\}$ NOE NMR studies. Using the DNA-binding domain as a structural reference, we show that the protein backbone of the N-terminal activation domain undergoes rapid, large-amplitude motions and is therefore unstructured. Difference CD data also show that the N-terminal activation domain remains random-coil, even in the presence of DNA. Implications for a "polypeptide lasso" model of transcriptional activation are discussed.

Keywords: activation domains; dynamics; heat shock factor; NMR; unstructured

Transcription factors orchestrate the regulated production of key proteins in eukaryotic cells during development and in response to extracellular stimuli. Typical transcription factors consist of a DNA-binding domain, an oligomerization domain, and one or more activation domains (Tjian & Maniatis, 1994). The structural basis for the function of the DNA-binding and oligomerization domains have been well characterized (Nelson, 1995); however, similar information for the activation domains remains scarce (Triezenberg, 1995). Many transcriptional activation domains fall into one of three categories on the basis of their predominant amino acid compositions: acidic, proline-rich, or glutamine-rich (Mitchell & Tjian, 1989). Fusion proteins produced by joining these domains with heterologous DNA-binding domains have exhibited transcriptional activation (Tjian & Maniatis, 1994). Several recent studies conducted on the "acidic" activation domains of Vmw65 protein of herpes simplex virus (Donaldson & Capone, 1992; O'Hare & Williams, 1992), NF-κB p65 (Schmitz et al., 1994), and the τ1 core of the human glucocorticoid receptor (Dahlman-Wright et al., 1995) concluded that these domains lacked well-defined structure. These results were based essentially on negative information: a lack of long-range NOEs and poor proton chemical shift dispersion in ^1H NMR spectra.

In this report, we use information on the dynamics of the protein backbone obtained from $^{15}N\{^1H\}$ heteronuclear NOE, along with CD experiments, to characterize the structures of the N-terminal transcriptional activation domains of the heat shock transcription factor (HSF) from two species of yeast. HSF is an inducible transcriptional activator that regulates the expression of the heat shock proteins when eukaryotic cells are exposed to elevated temperatures or other environmental insults (Lis & Wu, 1993; Morimoto, 1993). HSF contains a DNA-binding domain, a trimerization domain, and one or two transcriptional activation domains. In yeast, HSF is constitutively bound at heat shock elements containing the sequence 5′-nGAAn-3′ (Amin et al., 1988; Xiao & Lis, 1988) and acts as a transcription factor even under nonstressed conditions (Jakobsen & Pelham, 1988; Sorger & Pelham, 1988; Wiederrecht et al., 1988; Gross et al., 1990; Jakobsen & Pelham, 1991; Chen et al., 1993). Functional HSF is required for yeast viability at all temperatures (Sorger & Pelham, 1988; Wiederrecht et al., 1988; Gallo et al., 1993). Heat shock or stress conditions effect a higher level of transcriptional activation, which is mediated through mechanisms that are currently being investigated (Nieto-Sotelo et al., 1990; Sorger, 1990; Gallo et al., 1993).

Biochemical and genetic experiments have been used to map the transcriptional activity of yeast HSF to distinct regions of

Reprint requests to: David E. Wemmer, Department of Chemistry, University of California, Berkeley, California 94720; e-mail: dewemmer@lbl.gov.

*The first two authors contributed equally to this work.

the N- and C-termini (Nieto-Sotelo et al., 1990; Sorger, 1990; Jakobsen & Pelham, 1991) (Fig. 1). These activation domains could not be grouped into any known classes by their amino acid content, nor could they be categorized into any known folding motifs by structure prediction programs. The C-terminal activation domains from *Kluyveromyces lactis* and *Saccharomyces cerevisiae* exhibit high levels of constitutive activity when fused to a heterologous DNA-binding domain (Nieto-Sotelo et al., 1990; Sorger, 1990; Jakobsen & Pelham, 1991). The intact C-terminal activation domains are too large (300+ residues) for practical NMR studies. Although the residues most involved in activation in the C-terminal domain have been mapped to a much smaller segment (amino acids 592–623), its activity is clearly modulated by other regions of the protein (Chen et al., 1993), making this segment a questionable target for structural studies.

Deletions of the entire C-terminal region result in a functional HSF at physiological temperatures (Nieto-Sotelo et al., 1990; Sorger, 1990); this implies that the N-terminal activation domain is sufficient for the required constitutive level of transcriptional activity. In fact, the N-terminal activation domains from both yeast strains also show a low level of transcriptional activity when fused to a heterologous DNA-binding domain (Nieto-Sotelo et al., 1990; Sorger, 1990). The smaller size of the N-terminal activation domain (170–190 residues) makes it suitable for study by NMR. In addition, the N-terminal region is flanked by the well-studied and stable DNA-binding domain (Damberger et al., 1994; Harrison et al., 1994).

Using two constructs that contained both the N-terminal activation and the DNA-binding domains from *K. lactis* and *S. cerevisiae*, we were able to compare the structure and dynamics of the N-terminal activation domain against an internal structural control, the DNA-binding domain. $^{15}N\{^1H\}$ heteronuclear single quantum coherence (HSQC) spectra showed that the chemical shift dispersion of the ^1H resonances of the activation domains was poor, typical of that observed in denatured proteins (Neri et al., 1992; Shortle & Abeygunawardana, 1993). In addition, we used two-dimensional heteronuclear $^{15}N\{^1H\}$ NOE NMR and measurements of ^{15}N relaxation parameters to show that the N-terminal activation domains from both yeast strains have a high degree of flexibility, which is consistent with an unstructured state in solution. The results are particularly compelling because they offer the first positive evidence for a dynamically disordered transcriptional activation domain.

Results

Dynamics probed by heteronuclear relaxation

Heteronuclear relaxation studies provide a residue-specific probe of the conformational dynamics of proteins (Kay et al., 1989; Palmer, 1993). Current NMR approaches couple dynamic measurements with structural information to gain insight into protein function. Typical heteronuclear relaxation studies result in the determination of order parameters for backbone amide N-Hs (Clore et al., 1990; Barbato et al., 1992; Redfield et al., 1992; Cheng et al., 1993; Farrow et al., 1994; Zink et al., 1994). Order parameters, which define the amplitude and range of local molecular motions, are obtained using data acquired from a suite of three heteronuclear relaxation NMR experiments: ^{15}N T1 longitudinal relaxation, ^{15}N T2 transverse relaxation, and heteronuclear $^{15}N\{^1H\}$ NOE. In a recent example, Bax and coworkers applied such studies to support the "flexible tether" model for calmodulin function (Barbato et al., 1992).

The heteronuclear $^{15}N\{^1H\}$ NOE provides a qualitative assessment of the mobility of N-H bond vectors for individual residues. It is sensitive to both the overall tumbling time of the protein (τ_m) and fast internal motions. Maximal NOE values (+0.83) occur in the slow-tumbling limit ($\omega_N \tau_m \gg 1$), indicating the N-H bond vectors reorient with the overall tumbling of the molecule. Minimal NOE values (−3.5; assuming isotropic rotation, ^{15}N resonance frequency of 60.8 MHz, and 1.02 Å ^{15}N-^1H bond length) occur in the fast-tumbling limit ($\omega_N \tau_m \ll 1$) and are indicative of rapid, large-amplitude motions with respect to the overall tumbling of the molecule (Palmer, 1993).

Although heteronuclear NOE data alone are not sufficient for fully quantifying the dynamic behavior of molecules, they can be used as a probe for assessing the structural state of a protein. In one recent example, heteronuclear NOE data were particularly enlightening when comparing the backbone flexibilities between the folded and unfolded states of the drkN SH3 domain (Farrow et al., 1995). The data showed dramatic differences in dynamic behavior between the folded and unfolded states: the structured, folded state exhibited positive NOEs, and the unfolded state showed primarily negative NOEs.

Dynamic variations within the K. lactis HSF DNA-binding domain

Most of the cross peaks in the $^{15}N\{^1H\}$ HSQC spectra of the *K. lactis* HSF DNA-binding domain are well resolved. The spectrum for the heteronuclear NOE version of this experiment is shown in Figure 2A. Assignments were based on the previous NMR study (Damberger et al., 1994). Calculated NOE values, shown in Figure 3A, have an average of 0.67 (0.81 average in regions of secondary structure). The heteronuclear NOEs for the structured regions of the DNA-binding domain are positive and near the slow tumbling limit, as opposed to the C-terminal residues His-91 and Ala-92, which show negative NOEs indicative of rapid, large-amplitude motions. Areas of intermediate NOEs reflect an increase in dynamic flexibility. The best example of this can be seen for the L1 loop (residues 69–83). Previous structural work had suggested that there was higher mobility in this loop region, based on the lack of long-range contacts in the NMR structure (Damberger et al., 1994) and the lack of electron density for residues 76–79 in the crystal structure (Harrison et al., 1994). The heteronuclear NOE gives direct evidence that this region has a higher degree of dynamic flexibility with respect to the overall tumbling of the protein.

Structural assessment of the S. cerevisiae HSF DNA-binding domain

There are no published structures of the DNA-binding domain of *S. cerevisiae*; however, its high sequence homology (73%

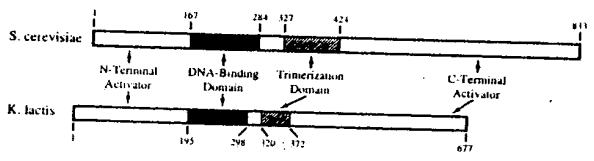

Fig. 1. Schematic representations of *K. lactis* and *S. cerevisiae* HSF.

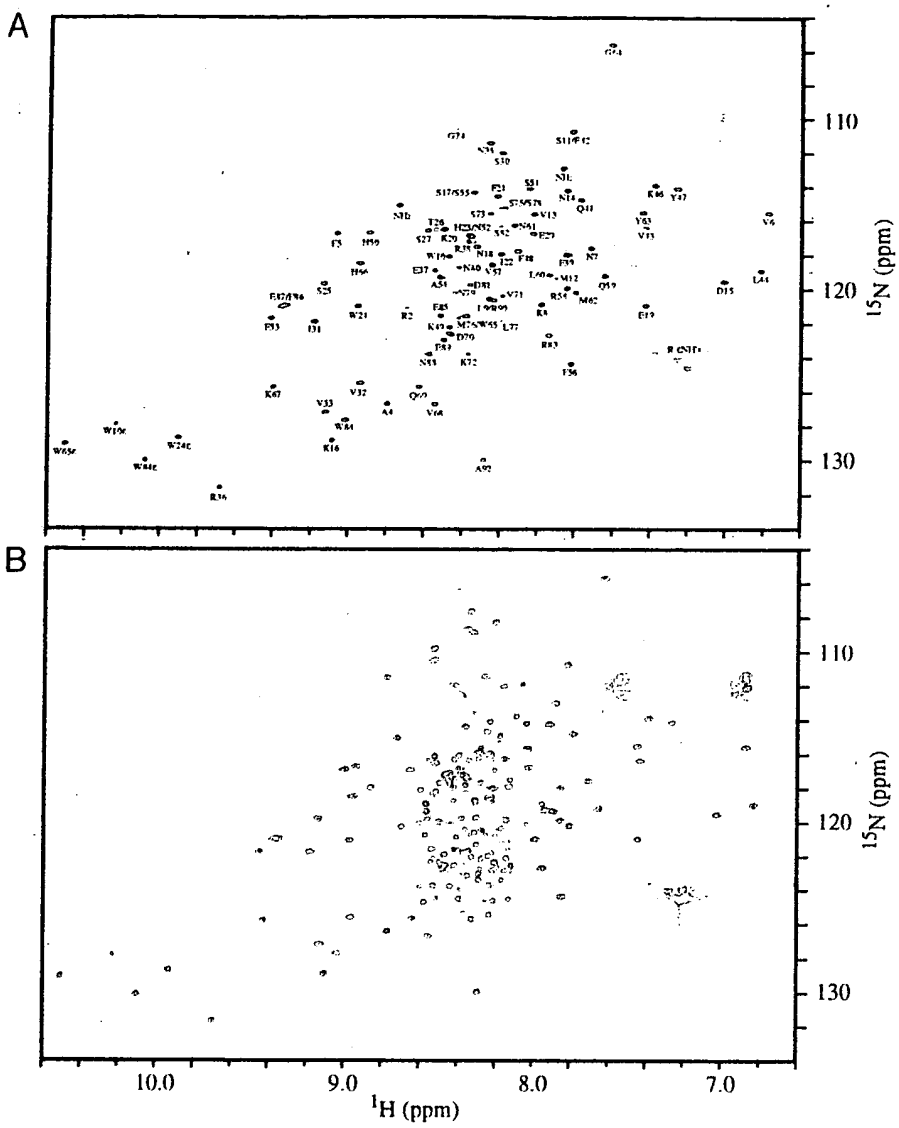

Fig. 2. Heteronuclear $^{15}N\{^1H\}$ NOE spectra for *K. lactis* HSF constructs. Positive intensity peaks are denoted in black; negative intensity peaks are denoted in red. **A:** DNA-binding domain construct. Cross peaks are labeled with their amino acid assignments (Damberger et al., 1994). **B:** Construct of the N-terminal activation domain plus the DNA-binding domain.

identity) (Devereux, 1991) and identical DNA-binding specificity to the DNA-binding domain of *K. lactis* HSF (Jakobsen & Pelham, 1991) strongly suggest that it has the same fold (Hubl et al., 1994). Indications that the DNA-binding domain of *S. cerevisiae* HSF is structured could be seen from the well-dispersed nature of the $^{15}N\{^1H\}$ HSQC spectra of this domain (data not shown) and from the similarity of the secondary structure to that of the *K. lactis* DNA-binding domain, as evidenced by far-UV CD experiments (data not shown). The presence of structure in this domain was verified with heteronuclear NOE experiments. Most of the calculated NOEs are positive and fall within the slow tumbling limit (data not shown). The average NOE value was 0.63 for the resolved peaks, just as in the *K. lactis* equivalent. Because of the well-behaved nature of this domain, we were able to use it as an internal structural control for analysis of the N-terminal activation domain of *S. cerevisiae* HSF.

N-terminal activation domain of yeast HSF is unstructured

The heteronuclear NOE spectrum of the *K. lactis* HSF N-terminal activation domain plus DNA-binding domain is shown in Figure 2B. The similarity of the peaks in Figure 2A with the positive peaks in Figure 2B facilitated the transfer of peak assignments. Of the 85 N-H signals previously assigned (Damberger et al., 1994), we were able to clearly identify 75 in our spectra. The respective NOE values for these peaks (Fig. 3B) correspond very well with the data for the DNA-binding domain alone (Fig. 3A). It appears that the backbone dynamics in most of the DNA-binding domain are not affected by the presence of the N-terminal activation domain.

The remaining peaks in Figure 2B, all negative in intensity, can be attributed to the N-terminal activation domain. These ad-

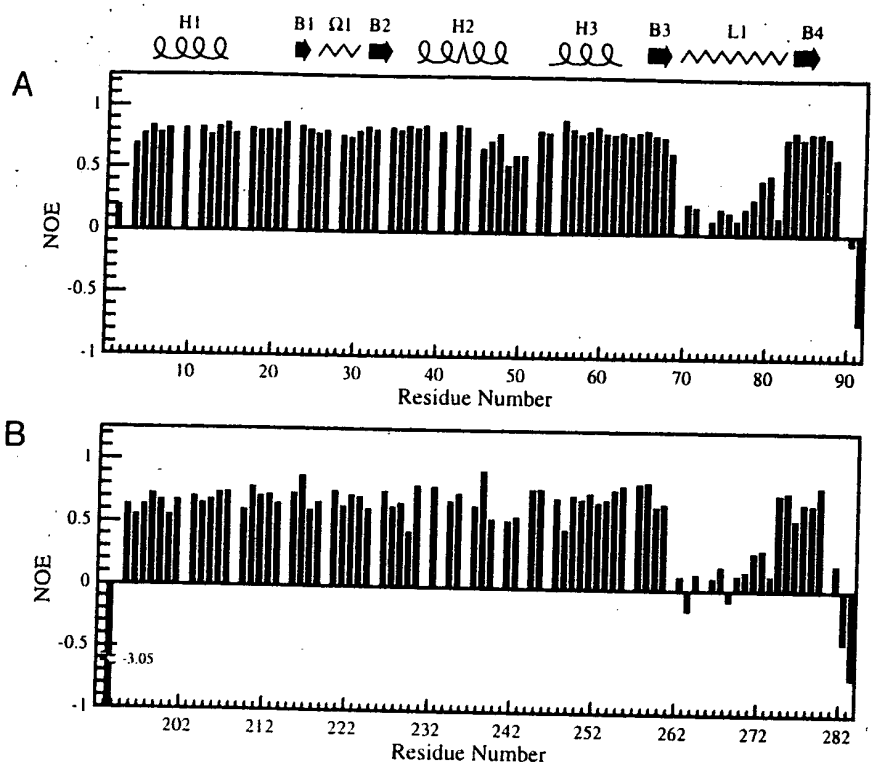

Fig. 3. Representations of the NOE data for the DNA-binding domains of the *K. lactis* HSF constructs. **A:** Assigned residues for the DNA-binding domain construct (1–92). **B:** Assigned DNA-binding domain residues (193–284) for the construct of the N-terminal activation domain plus the DNA-binding domain.

ditional peaks have poor shift dispersion in the proton dimension, as is the case for unstructured proteins. Integrating the volumes of the signals in the spectra could account for approximately 183 of the possible 192 residues; however, only 103 peaks were sufficiently resolved to calculate NOEs. These NOE values (Fig. 4) are sorted by proton chemical shift because there are no sequence-specific assignments currently available for this domain. Negative values indicate that all of the residues in this region exhibit a high degree of rapid internal motion, typical for an unstructured protein. Rapid motions of the activation domain also affect the dynamics of the N-terminus of the DNA-binding domain. Arg 2 in the DNA-binding domain alone showed a small positive NOE (Fig. 3A), but the presence of the highly dynamic activation domain in the fusion construct of the N-terminal activation domain plus the DNA-binding domain seems to drive the corresponding residue (Arg 194) into a rapid motion regime (Fig. 3B).

Identical experiments were performed on the comparable construct for *S. cerevisiae* HSF. Despite the lack of residue-specific assignments for the DNA-binding domain, we were still able to identify the corresponding peaks in the full-length construct by comparison with spectra of the DNA-binding domain alone.

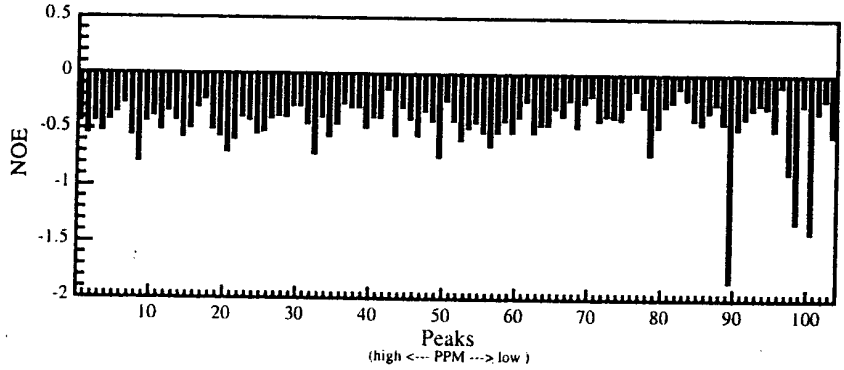

Fig. 4. NOE data for the activation domain peaks of the construct of the N-terminal activation domain plus the DNA-binding domain. Histogram sorted by proton chemical shift.

Again, the calculated NOEs fell predominately in the slow tumbling regime. The remaining signals, attributable to the N-terminal activation domain, all had negative NOEs indicative of rapid, large-amplitude motions. This shows that the N-terminal activation domain of *S. cerevisiae* also behaves as an unstructured protein.

To confirm that the N-terminal activation domains were not unstructured because of denaturation at low pH, further NMR spectroscopic studies were performed at pH 5.75 (data not shown). Comparison of the chemical shift distribution in spectra acquired at higher pH with those at lower pH indicated no change in the structural state of the constructs. The DNA-binding domains remained folded, whereas the activation domains persisted in an unstructured state.

Far-UV CD spectroscopy, performed at pH 7.0, also indicate that the N-terminal activation domains of yeast HSF are unstructured. Figure 5A shows CD spectra for the DNA-binding domain, and for the N-terminal activation domain plus the DNA-binding domain, as well as the difference between the two. The difference spectrum is characteristic of a random coil. We also examined the possibility that the N-terminal activation domain might behave differently when the DNA-binding domain is bound to a DNA-binding site (Fig. 5B). An examination of this CD difference spectrum indicates that binding of the DNA-binding domain to the DNA does not induce any structural changes in the N-terminal activation domain. Identical results were obtained for the *K. lactis* constructs (not shown).

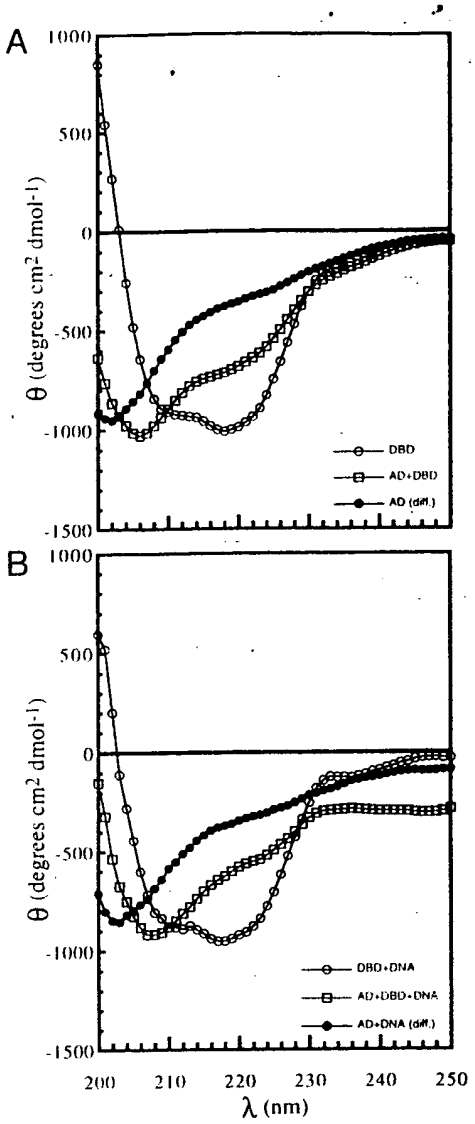

Fig. 5. Far-UV CD spectra of *S. cerevisiae* HSF constructs. **A:** Protein only. **B:** Protein and DNA. DBD, DNA-binding domain construct; AD + DBD, construct of the N-terminal activation domain plus the DNA-binding domain; AD (diff.), activation domain difference spectra.

Discussion

Progress is being made in the effort to understand the mechanism of transcriptional activation. Structural studies of proteins often reveal important aspects of their function and are a powerful complement to biochemical and genetic studies. Heteronuclear relaxation NMR spectroscopy can be used to characterize the behavior of unstructured or partially structured protein states in solution. Using heteronuclear NOE experiments, we have been able to show that the N-terminal activation domain of yeast HSF undergoes rapid local motions and is unstructured in solution. It is possible that the structural state of the N-terminal activation domain might be affected by the presence of the deleted C-terminal regions. However, this seems unlikely because the N-terminal activation domain is functional when fused to heterologous DNA-binding domains (Nieto-Sotelo et al., 1990), and C-terminal deletion constructs are sufficient for cell viability (Nieto-Sotelo et al., 1990; Sorger, 1990).

There have been several previous studies of functional fragments of acidic activation domains, varying in both size and source. The consensus of these studies is that these peptides alone in solution have little secondary structure, although data suggest that either α helices (Donaldson & Capone, 1992; Schmitz et al., 1994; Dahlman-Wright et al., 1995) or β sheets (Leuther et al., 1993; Vanhoy et al., 1993) can be induced under specific solvent conditions. It has then been concluded that these conformations might be important in the function of the domains, although there is no direct evidence for this. Models have also been presented that suggest that activation domains interact through amphipathic helices (Giniger & Ptashne, 1987), but these models have limited support from either genetic or biochemical data. It has been clearly shown, however, that specific patterns of acidic and hydrophobic residues are required for activity of acidic activation domains (Cress & Triezenberg, 1991; Regier et al., 1993), although this does not seem to hold for glutamine- or proline-rich domains (Gerber et al., 1994). The studies presented here are the first on a general sequence activation domain (not falling into acidic, proline-rich, or glutamine-rich classes), and also the first on a complete activation domain. The NMR data clearly show that this domain is highly disordered, with internal motions on a very rapid time scale. We have reached somewhat similar conclusions about an activation domain fragment from c-jun (unpubl. data), which again does not fall into any of the classes noted above. These data, taken together with previous studies of acidic activation domains, argue strongly that the disorder observed is in fact a general property of activation domains of all classes.

It is not obvious how an unstructured activation domain might interact with the components of the general transcriptional machinery. An "induced fit" model proposes that some proteins

may function by acquiring structure in the presence of target factors (Koshland, 1958). Induced fit structural changes have been found in systems such as b-ZIP binding to DNA (Talanian et al., 1990), calmodulin binding to target proteins (Ikura et al., 1992), and staphylococcal nuclease binding to inhibitor (Hynes & Fox, 1991). However, these examples are of proteins that already possess some degree of structure in their "apo" forms. The structural changes induced in these proteins by their interacting partners are relatively small and are localized around the substrates. No conclusions can be drawn about the structure of activation domains in their activated complexes until further data is gathered, but it would be surprising to find that systems as large and dynamically disordered as activation domains could become completely folded via an induced fit mechanism.

With respect to this, we suggest that activation domains only serve as "polypeptide lassos" that rein in and increase the local concentrations of the factors needed for activated transcription. Evidence for this type of role can be drawn from reports indicating that activation domains interact with several different factors of the general transcriptional machinery (Tjian & Maniatis, 1994). The localization of these factors could stimulate the formation of the transcriptional complex and thereby initiate transcription. The exact nature of this recruiting process is still not clear, but some insight can be gained from examining the biochemical data on the interaction of acidic domains with the general transcription factor TFIIB. Recently published results show that acidic activators interact with the C-terminal region of TFIIB to recruit this factor into the preinitiation complex (Roberts & Green, 1994). However, the protein-protein interaction that mediates this and similar events are unlikely to be highly specific in nature. Experiments using heterologous transcription factors formed by a fusion of GAL4 or GCN4 DNA-binding domains to random polypeptides coded by fragments of genomic *Escherichia coli* DNA found the polypeptides that provided the highest levels of enhanced transcription to be rich in acidic side chains but to possess no other obvious sequence similarity (Ma & Ptashne, 1987). This reinforces the notion that nonspecific interactions can play a large role in the process of activated transcription, a generalization of the "acid blobs and negative noodles" idea of Sigler (1988). It would then not be too surprising if other activation domains recruited their factors in a similar, nonspecific fashion. Once needed factors are gathered, their close proximity to each other could stimulate the assembly of the active transcription complex. Similar proposals have been outlined by Frankel and Kim (1991).

Thus, it may not be necessary for activation domains to form any well-defined and unique structure. The idea of activation domains functioning as "polypeptide lassos" better supports the impression that a small number of transcription factors are able to regulate the production of thousands of proteins (Tjian & Maniatis, 1994). Their lack of a well-defined and unique structure could make activation domains more versatile in their abilities to recruit the various factors necessary for formation of the transcriptional complex.

Materials and methods

Cloning

The HSF activation domain plus DNA-binding domain constructs were cloned from plasmids developed previously (Flick et al., 1994; Harrison et al., 1994). The plasmid pHN200, which contains the coding sequence for the *K. lactis* HSF, was modified by site-directed mutagenesis with two oligomers to insert an *Nde*I site at the N-terminus and an *Sph*I site at amino acid 284. This restriction fragment was subcloned into *Nde*I/*Sph*I digested pHN104, a T7-driven expression vector. The *S. cerevisiae* construct was made in a similar manner from plasmid pHF153, which codes for the first 259 amino acids of HSF. It already contained the correct *Sph*I site at residue 259. An *Nde*I site was introduced at the N-terminus with site-directed mutagenesis, and again the *Nde*I/*Sph*I fragment was subcloned into pHN104 for expression.

Sample preparation

Plasmid constructs were transformed into *E. coli* (strain BL21 (DE3)/pACYC) (Flick et al., 1994; Harrison et al., 1994). Uniform ^{15}N-labeled protein was obtained by growing the bacteria on M9 minimal medium with [^{15}N]NH_4Cl (CIL) as the sole nitrogen source. Induction occurred at OD_{594} = ~0.6 with 1.2 g IPTG/1L, and the cells were harvested 4 h later by centrifugation at 4,000 rpm, 4 °C, in a Sorvall GS3 rotor. Protein purification was accomplished by a simplified version of a previously reported method (Harrison, 1994). Cells were resuspended in 10 mL/g cells isotonic wash buffer, spun down, resuspended in 1 M NaCl lysis buffer, and lysed by sonication for 6 min in a dry ice/ethanol bath. A high-speed spin was used to get rid of cellular debris, and the resultant supernatant was diluted to 100 mM NaCl and loaded onto a pre-equilibrated heparin column. Prior to elution, the column was washed with a low-salt buffer. The desired protein was eluted with 500 mM NaCl. PAGE (15%) was used to assess the purity of fractions collected. Electrospray-ionization mass spectroscopy confirmed >98% ^{15}N incorporation. Fractions containing >95% of the desired protein were dialyzed into NMR buffer to yield final 0.5-mL samples of ~3–5 mM protein in 10 mM KH_2PO_4, 90% H_2O/10% D_2O, at pH 3.4 or pH 5.75.

NMR spectroscopy

^{15}N[^{1}H] heteronuclear NOE spectra were measured on a Bruker AMX-600 instrument as described (Barbato et al., 1992). Spectra were initially acquired at pH 3.4 to minimize proton exchange and maximize the signal-to-noise ratio. The 8-k data points were collected in the $t2$ dimension to increase resolution. Data were processed with the NMRPipe suite of programs (Delaglio, 1993), and peaks picked with the CAPP/PIPP suite of analysis programs (Garrett et al., 1991). Spectra at pH 5.75 were collected with the PEP-Z-HSQC experiment (Akke et al., 1994). Data were processed with the Felix processing package (Biosym Technologies).

CD spectroscopy

Far-UV CD spectroscopy was performed on an AVIV Model 62DS equipped with a temperature-controlled cell holder and connected to an IBM-compatible workstation for data analysis. Samples were prepared as 1 mg/mL solutions in 25 mM sodium phosphate buffer, pH 7.0, at 25 °C. Concentrations (2:1) of DNA(5′-CCGGTGAATTTTCTTGAATGGCC-3′):protein were used for the protein/DNA experiments. Difference spectra were

obtained using Microsoft Excel. A three-point moving average was used for data smoothing.

Acknowledgments

We gratefully thank Karen Flick for assistance with the CD measurements and HPLC, Dr. David King for performing ESI mass spectral analysis, and Prof. Susan Marqusee for use of her CD spectrometer. This work has been supported by an award from the PEW Scholars Program in Biomedical Sciences (to H.C.M.N.), the Director, Office of Energy Research, Office of Biological & Environmental Research, General Life Science Division of the U.S. Department of Energy under Contract No. DE-AC03-76F00098 (to D.E.W.), and NIH Traineeship GM08295 (to C.W.L.). Instrumentation grants were provided by the U.S. Department of Energy, DE-FG05-86ER75281 and the NSF, DMB86-09305 and BBS 87-20134 (to D.E.W.).

References

Akke M, Carr PA, Palmer AG. 1994. Heteronuclear-correlation NMR spectroscopy with simultaneous isotope filtration, quadrature detection, and sensitivity enhancement using z rotations. *J Magn Reson B 104*:298-302.

Amin J, Ananthan J, Voellmy R. 1988. Key features of heat shock regulatory elements. *J Mol Biol 8*:3761-3769.

Barbato G, Ikura M, Kay LE, Pastor RW, Bax A. 1992. Backbone dynamics of calmodulin studied by ^{15}N relaxation using inverse detected two-dimensional NMR spectroscopy: The central helix is flexible. *Biochemistry 31*:5269-5278.

Chen Y, Barlev NA, Westergaard O, Jakobsen BK. 1993. Identification of the C-terminal activator domain in yeast heat shock factor: Independent control of transient and sustained transcriptional activity. *EMBO J 12*: 5007-5018.

Cheng JW, Lepre CA, Chambers SP, Fulghum JR, Thomson JA, Moore JM. 1993. ^{15}N NMR relaxation studies of the FK506 binding protein: Backbone dynamics of the uncomplexed receptor. *Biochemistry 32*:9000-9010.

Clore GM, Driscoll PC, Wingfield PT, Gronenborn AM. 1990. Analysis of the backbone dynamics of interleukin-1β using two-dimensional inverse detected heteronuclear ^{15}N-^{1}H NMR spectroscopy. *Biochemistry 29*:7387-7401.

Cress WD, Triezenberg SJ. 1991. Critical structural elements of the VP16 transcriptional activation domain. *Science 251*:87-90.

Dahlman-Wright K, Baumann H, McEwan IJ, Almlof T, Wright APH, Gustafsson JA, Hard T. 1995. Structural characterization of a minimal functional transactivation domain from the human glucocorticoid receptor. *Proc Natl Acad Sci USA 92*:1699-1703.

Damberger FF, Pelton JG, Harrison CJ, Nelson HCM, Wemmer DE. 1994. Solution structure of the DNA-binding domain of the heat shock transcription factor determined by multidimensional heteronuclear magnetic resonance spectroscopy. *Protein Sci 3*:1806-1821.

Delaglio F. 1993. *NMRPipe system of software*. Bethesda, Maryland: National Institutes of Health.

Devereux J. 1991. *Program manual for the GCG package, version 7*. Madison, Wisconsin: Genetics Computer Group, Inc.

Donaldson L, Capone JP. 1992. Purification and characterization of the carboxyl-terminal transactivation domain of Vmw65 from herpes simplex virus type I. *J Biol Chem 267*:1411-1414.

Farrow NA, Muhandiram R, Singer AU, Pascal SM, Kay CM, Gish G, Shoelson SE, Pawson T, Forman-Kay JD, Kay LE. 1994. Backbone dynamics of a free and a phosphopeptide-complexed Src homology 2 domain studied by ^{15}N NMR relaxation. *Biochemistry 33*:5984-6003.

Farrow NA, Zhang O, Forman-Kay JD, Kay LE. 1995. Comparison of the backbone dynamics of a folded and an unfolded SH3 domain existing in equilibrium in aqueous buffer. *Biochemistry 34*:868-878.

Flick K, Gonzalez LJ, Harrison CJ, Nelson HCM. 1994. Yeast heat shock factor contains a flexible linker between the DNA-binding and trimerization domains; Implications for DNA-binding by trimeric proteins. *J Biol Chem 269*:12475-12481.

Frankel AD, Kim PS. 1991. Modular structure of transcription factors: Implications for gene regulation. *Cell 65*:717-719.

Gallo GJ, Prentice H, Kingston RE. 1993. Heat shock factor is required for growth at normal temperatures in the fission yeast *Schizosaccharomyces pombe*. *Mol Cell Biol 13*:749-761.

Garrett DS, Powers R, Gronenborn AM, Clore GMA. 1991. A common sense approach to peak picking in two- three- and four- dimensional spectra using automatic computer analysis of contour diagrams. *J Magn Reson 95*:214-220.

Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, Schaffner W. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science 263*:808-811.

Giniger E, Ptashne M. 1987. Transcription in yeast activated by a putative amphipathic alpha helix linked to a DNA binding unit. *Nature 330*: 670-672.

Gross DS, English KE, Collins KW, Lee SW. 1990. Genomic footprinting of the yeast HSP82 promoter reveals marked distortion of the DNA helix and constitutive occupancy of heat shock and TATA elements. *J Mol Biol 216*:611-631.

Harrison CJ. 1994. The structure of the DNA-binding domain from *Kluyveromyces lactis* heat shock transcription factor [thesis]. Berkeley, California: University of California.

Harrison CJ, Bohm AA, Nelson HCM. 1994. Crystal structure of the DNA binding domain of the heat shock transcription factor. *Science 263*: 224-227.

Hubl ST, Owens JC, Nelson HCM. 1994. Mutational analysis of the DNA-binding domain of yeast heat shock transcription factor. *Struct Biol 1*:615-620.

Hynes TR, Fox RO. 1991. The crystal structure of staphylococcal nuclease refined at 1.7 Å resolution. *Proteins Struct Funct Genet 10*:92-105.

Ikura M, Clore GM, Gronenborn AM, Zhu G, Klee CB, Bax A. 1992. Solution structure of a calmodulin-target peptide complex by multidimensional NMR. *Science 256*:632-638.

Jakobsen BK, Pelham HRB. 1988. Constitutive binding of yeast heat shock factor to DNA in vivo. *Mol Cell Biol 8*:5040-5042.

Jakobsen BK, Pelham HRB. 1991. A conserved heptapeptide restrains the activity of the yeast heat shock transcription factor. *EMBO J 10*:369-375.

Kay LE, Torchia DA, Bax A. 1989. Backbone dynamics of proteins as studied by ^{15}N inverse detected heteronuclear NMR spectroscopy: Application to staphylococcal nuclease. *Biochemistry 28*:8972-8979.

Koshland DE. 1958. Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci USA 44*:98-104.

Leuther KK, Salmeron JM, Johnston SA. 1993. Genetic evidence that an activation domain of Gal4 does not require acidity and may form a β-sheet. *Cell 72*:587-594.

Lis J, Wu C. 1993. Protein traffic on the heat shock promoter: Parking, stalling, and trucking along. *Cell 74*:1-4.

Ma J, Ptashne M. 1987. A new class of yeast transcriptional activators. *Cell 51*:113-119.

Mitchell PJ, Tjian R. 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science 245*:371-378.

Morimoto RI. 1993. Cells in stress: Transcriptional activation of heat shock genes. *Science 259*:1409-1410.

Nelson HCM. 1995. Structure and function of DNA-binding proteins. *Curr Opin Genet Dev 5*:180-189.

Neri D, Wider G, Wüthrich K. 1992. Complete ^{15}N and ^{1}H NMR assignments for the amino-terminal domain of the phage 434 repressor in the urea-unfolded form. *Proc Natl Acad Sci USA 89*:4397-4401.

Nieto-Sotelo J, Wiederrecht G, Okuda A, Parker CS. 1990. The yeast heat shock transcription factor contains a transcriptional activation domain whose activity is repressed under nonshock conditions. *Cell 62*:807-817.

O'Hare P, Williams G. 1992. Structural studies of the acidic transactivation domain of the Vmw65 protein of herpes simplex virus using ^{1}H-NMR. *Biochemistry 31*:4150-4156.

Palmer AG. 1993. Dynamic properties of proteins from NMR spectroscopy. *Curr Opin Biotechnol 4*:385-391.

Redfield C, Boyd J, Smith LJ, Smith RAG, Dobson CM. 1992. Loop mobility in a four-helix-bundle protein: ^{15}N NMR relaxation measurements on human interleukin-4. *Biochemistry 31*:10431-10437.

Regier JL, Shen F, Triezenberg SJ. 1993. Pattern of aromatic and hydrophobic amino acids critical for one of two subdomains of the VP16 transcriptional activator. *Proc Natl Acad Sci USA 90*:883-887.

Roberts SGE, Green MR. 1994. Activator-induced conformational change in general transcription factor TFIIB. *Nature 371*:717-720.

Schmitz ML, Silva MAS, Altmann H, Czisch M, Holak TA, Baeuerle PA. 1994. Structural and functional analysis of the NF-κB p65 C terminus. *J Biol Chem 269*:25613-25620.

Shortle D, Abeygunawardana C. 1993. NMR analysis of the residual structure in the denatured state of an unusual mutant of staphylococcal nuclease. *Structure 1*:121-134.

Sigler PB. 1988. Transcriptional activation. Acid blobs and negative noodles. *Nature 333*:210-212.

Sorger PK. 1990. Yeast heat shock factor contains separable transient and sustained response transcriptional activators. *Cell 62*:793-805.

Sorger PK, Pelham HRB. 1988. Yeast heat shock factor is an essential DNA-binding protein that exhibits temperature-dependent phosphorylation. *Cell 54*:855–864.

Talanian RV, McKnight CJ, Kim PS. 1990. Sequence-specific DNA binding by a short peptide dimer. *Science 249*:769–771.

Tjian R, Maniatis T. 1994. Transcriptional activation: A complex puzzle with few easy pieces. *Cell 77*:5–8.

Triezenberg SJ. 1995. Structure and function of transcriptional activation domains. *Curr Opin Genet Dev 5*:190–196.

Vanhoy M, Leuther KK, Kodadek T, Johnston SA. 1993. The acidic activation domains of the GCN4 and Gal4 proteins are not α-helical but form β-sheets. *Cell 72*:587–594.

Wiederrecht G, Seto D, Parker CS. 1988. Isolation of the gene encoding the *S. cerevisiae* heat shock transcription factor. *Cell 54*:841–853.

Xiao H, Lis JT. 1988. Germline transformation used to define key features of heat-shock response elements. *Science 239*:1139–1142.

Zink T, Ross A, Lüers K, Cieslar C, Rainer R, Holak TA. 1994. Structure and dynamics of the human granulocyte colony-stimulating factor determined by NMR spectroscopy. Loop mobility in a four-helix-bundle protein. *Biochemistry 33*:8453–8463.

To appear in the Proc. IEEE International Conference on Neural Networks, Huston, TX, June 1997

Identifying Disordered Regions in Proteins from Amino Acid Sequence[1]

P. Romero[2], Z. Obradović[2], C. Kissinger[4], J. E. Villafranca[4], and A. K. Dunker[3]

[2]School of Electrical Engineering and Computer Science
[3]Department of Biochemistry and Biophysics
Washington State University, Pullman, Washington, 99164-2752
and
[4]Agouron Pharmaceuticals, Inc.
3565 General Atomics Ct., San Diego, CA 92121-1221

Abstract

A rule-based and several neural network predictors are developed for identifying disordered regions in proteins. The rule-based predictor was suitable only for very long disordered regions, whereas the neural network predictors were developed separately for short-, medium-, and long-disordered regions (S-, M-, and LDRs, respectively). The out-of-sample prediction accuracies on a residue-by-residue basis ranged from 69 to 74% for the neural network predictors when applied to the same length class, but fell to 59 to 67% when applied to different length classes. Application of both the rule-based and LDR neural network predictors to large databases of protein sequences provide strong evidence that disordered regions are very common in nature. These results are consistent with our recent proposal that disordered regions are crucial for the evolution of molecular recognition.

1 Introduction

X-ray diffraction from protein crystals can yield the structure of these molecules. Many proteins are nonuniform, having both structured and disordered regions. When crystallized, the structured regions scatter x-rays coherently and so are observed. The disordered regions, however, fail to crystallize into fixed structures and so scatter x-rays incoherently. These disordered regions are therefore invisible in the resulting electron density maps [8].

We have so far identified more than 15 proteins containing disordered regions that become ordered upon formation of a molecular complex with a partner. Homology searches extend these to hundreds of examples. These include the following types of molecular interactions: enzyme/substrate, receptor/ligand, protein/protein, protein/RNA and protein/DNA. Thus, disorder-to-order transitions upon binding span the biological domain [4]. Schulz proposed that loss of disorder upon binding leads to a biologically desirable pair of features, namely: high specificity coupled with modest binding affinities [13]. Affinities that are too high result in essentially irreversible binding, which is unsuitable for most biological processes.

For proteins whose x-ray structures are known, the existence of disordered stretches can be identified directly by looking for amino acids that are missing from the electron density maps. However, x-ray structures are known for just a small fraction of the set of proteins with known amino acid sequences. Consequently, to estimate the occurance of proteins with disordered regions in nature, alternative information-based approaches have to be taken. The approach considered in this study is to develop predictors that can identify disordered regions in proteins. It is well-established that amino acid sequence determines protein 3-D structure [2]; here we assume that amino acid sequence determines lack of fixed 3-D structure (e.g. *disorder*) as well.

Design of a rule-based predictor is summarized in Section 2. Construction of a labeled data set of proteins with disordered regions and its use to develop neural network predictors is discussed in Section 3. Finally, both approaches were tested on several databases as reported in Section 4.

[1]Corresponding E-mail: zoran@eecs.wsu.edu

2 Rule-Based Approach

Studies on calcineurin (CaN) [9] sparked the present work; we noticed that the long disordered region (LDR) in this protein has a low content of aromatic amino acids (Trp, Tyr, Phe). Several other disordered regions were found to have this same characteristic. This makes structural sense because the side chains of aromatic amino acids have strong and specific interactions [3] and so would be expected to induce structure and inhibit disorder. Using CaN as the prototype, the average fraction of aromatic amino acids was calculated over a window of 31 amino acids surrounding each sequence position. The aromatic content dropped significantly in both of the longer unobserved regions in CaN (Fig 1).

The following prediction rule was developed from these observations: (a) for a given protein, the average content of aromatic residues is calculated throughout the amino acid sequence, as explained above; (b) if there is a contiguous region of more than 80 sequence positions with an average content of aromatic residues below 6.5%, the protein is predicted to have an LDR. This predictor was intended for LDRs like the one in CaN; because of the large window sizes for this predictor, it is not suitable for predicting M- or SDRs.

3 Neural Network Approach

The rule-based predictor discussed in the previous section was developed based on information from a single protein, CaN, which served as the prototype for our studies. An alternative is to design feedforward neural network predictor trained using the backpropagation learning algorithm [15]. This predictor requires construction of a larger set of examples of disordered regions (DRs) and determination of appropriate features, as discussed in this section.

3.1 Disordered Regions Labeled Data Sets

A search for proteins with invisible regions, which are presumed to be locally disordered, was performed on the Protein Data Bank (PDB) at the Brookhaven National Laboratory. This is a public domain archive of more than 4,600 experimentally determined three-dimensional structures of proteins.

Searching PDB for proteins with DRs is a non-trivial task since no standard format for reporting such findings is imposed. In addition, several other problems like complexes and repeated sequences further complicated this search. In this study, no effort was made to identify all proteins with DRs in the PDB. The main objective was to find a sufficiently large set of proteins with confirmed DRs as needed for the design of a neural network predictor.

The PDB search supplied a set of proteins each having at least one DR longer than seven residues. These proteins from PDB were supplemented by two additional proteins with DRs (CaN [9] and Bcl [12]). A histogram of the lengths of the DRs suggested a partition into short, medium and long labeled data sets, denoted as SDR (7-21 amino acids), MDR (22-44 amino acids), and LDR (45 or more amino acids), respectively.

3.2 Feature Selection

The LDR labeled data set was analyzed to identify a pool of properties that discriminate between structured and disordered regions. The exploratory analysis considered several attributes measured by averaging over windows of consecutive amino acids. Considered attributes included individual amino-acid compositions, flexibility [14], hydropathy [11] and hydrophobic moments [5].

In addition to the lack of aromatics (in this case just Tyr and Trp) mentioned above, low amounts of Cys and His and high amounts of Glu, Asp, Ser, and Lys were also found to be associated with disorder. Cys can make special covalent bounds, so its absence in disordered regions is reasonable. Glu, Asp, and Lys are charged; charge imbalance would be expected to contribute to disorder. Ser increases solubility and provides a flexible locus. Finally, disordered regions would be expected to be soluble and flexible in a manner consistent with the findings on hydrophathy and flexibility. Thus, overall, the identified attributes seem reasonable for promoting disorder.

A set of m attributes was identified through this analysis as being more discriminative. For each identified attribute, n different features were generated by computing this attribute for n different window sizes, yielding

an $m \times n$ matrix of features, where each row corresponds to a different attribute and each column to a different window size.

A formal procedure was used to select the most appropriate feature from each row of this matrix. The method used here is an adaptation of the sequential forward search feature selection technique with the minimal error probability selection criterion [7]. A quadratic Gaussian classifier using different covariance matrices for each class was used to calculate minimal error probability during the search.

The standard sequential forward search selection technique is a greedy algorithm that begins with an empty feature set and adds features to it one at a time. The first feature added is the one deemed to be the best according to the selection criteria. The next feature added is the one which results in the largest improvement when considered in conjunction with the first feature. Similarly, the i-th feature added is the one that results in the largest improvement when considered in conjunction with the previous $i-1$ features.

In the method used here, when the i-th feature is added to the selected features set, its corresponding row is removed from the matrix and the search continues on the reduced $(m-i) \times n$ matrix. This prevents the same attribute from being selected with more than one window size; the resulting selected feature set contains the most appropriate window size for each attribute. A set of examples for neural network training was constructed from the LDR labeled data set using the m selected features. For convenience the same features set was used when training on examples from the SDR and MDR labeled data sets.

4 Results

The SDR labeled data set contained 38 disordered segments from 34 proteins with 411 disordered amino acids and 11,050 total amino acids; MDR set contained 22 disordered segments from 20 proteins with 464 disordered amino acids and 4,764 total amino acids; and finally, LDR set contained 7 regions from 7 proteins with 465 amino acids and 2,069 total amino acids. The result of the exploratory analysis on 24 considered attributes was the selection of 10, shown in Table 1. Shown here are the most appropriate windows for each of these attributes obtained through the feature selection process discussed in Section 3.2 by exploring odd-numbered values ranging from 9 to 21.

4.1 Prediction Accuracy Estimates

The rule-based predictor was designed using the CaN knowledge. When tested on a residue-by-residue prediction on the remaining 6 LDR proteins, it achieved 70% success rate. This result is surprisingly good for the simplicity of the rule and suggests that lack of aromatic amino acids is a strong determinant for the development of LDRs.

Balanced sets of the 10 dimensional feature values corresponding to the unobserved and observed amino acids were constructed from the S-, M-, and LDR labeled data sets. These feature sets were each randomly partitioned into 5 disjoint balanced subsets from observed and unobserved amino acids. A neural network architecture was determined through limited experimentation and a machine with 10 inputs, one hidden layer with 6 units, and a single output unit was used for in depth testing. 5-cross validation experiments starting from 3 different random initializations of neural network parameters were performed, resulting in a total of 15 runs each for the S-, M- and LDR labeled data sets (Table 2, rows a-e).

Averaging the results from the 5-cross validation experiments gave residue-by-residue prediction accuracies ranging from 66 to 74%. Lumping all length-classes together (ADR) and repeating the 5-cross validation experiment led to a drop in prediction accuracy, to about 60%. Finally, when testing each predictor on data sets of other length classes, the prediction accuracies dropped to 59-67%. The greatest drop was observed

Selected	composition								average	
Attribute	His	Glu	Lys	Ser	Asp	Cys	Trp	Tyr	Hydropathy	Flexibility
Window Size	9	9	9	9	13	21	21	21	9	15

Table 1: Selected Features

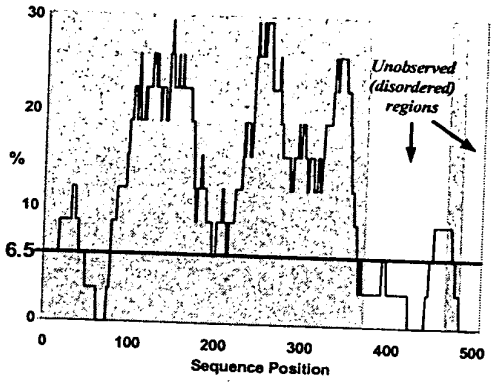

Figure 1: Fraction of aromatic residues on CaN

Test Set	Out of Sample Accuracy			
	SDR	MDR	LDR	ADR
a	71%	75%	70%	60%
b	71%	74%	77%	62%
c	69%	76%	77%	57%
d	67%	72%	72%	61%
e	66%	75%	70%	60%
Ave	69%	74%	73%	60%
SDR	-	63%	59%	-
MDR	63%	-	67%	-
LDR	59%	61%	-	-

Table 2: NN generalization results (5-cross validation averaged over 3 initializations)

for the LDR predictor applied to the SDR labeled data set or vice versa (both to 59%). The smallest drop was observed for the LDR predictor applied to the MDR labeled data set (67%).

4.2 Prediction of Disordered Regions in Proteins

The rule-based predictor produces binary outputs that are directly applicable for observability predictions for stretches of proteins due to the fact that the aromatic residues' content changes slowly when a window of 31 residues slides from one sequence position to the next. This is not true for the neural network predictor whose outputs are real numbers in (0,1) range that can vary significantly between adjacent positions. Consequently, neural network outputs are averaged over a window of 9 neighboring residues in order to smooth the signal as necessary for predicting disordered regions. Using a 0.5 threshold this signal is discretized to 0 or 1 corresponding to structured and disordered residue prediction, and the resulting binary signal is checked for regions of contiguous prediction of unobserved residues.

Estimating false positive error rates for the predictors is key to determining whether disordered regions are common. The presence of undetermined false negative error rates means only that our estimates represent lower bounds for the numbers of DRs. An estimate of false positive error rates was accomplished using the NRL_3D database, which contains the sequences of the observed parts of the proteins in PDB. Consequently, a perfect predictor should not accept any protein from NRL_3D.

The LDR predictor was applied to four data bases: NRL_3D (to estimate false positives), PDB (known to contain proteins with disordered regions) and also to two large sequence-only databases, SwissProt (SW) and the Protein Identification Resource (PIR). SW and PIR contain 59,031 and 89,926 sequences, respectively. Figure 2 and its accompanying Table 3 show the percentage of sequences predicted to have at least one predicted unobserved stretch longer than t. Due to the lower accuracy of the LDR predictor in the SDR region, data for $t < 20$ should be ignored.

The curve for the predicted DR rates in PDB is shifted to the right of the curve for NRL_3D. Examination of a few specific proteins indicates that this right-ward shift is due mostly to correct predictions.

From 20 to about 35 amino acids, the predicted rates of DRs in PDB, SW, and PIR are very similar. In contrast, the predicted DR rates in SW and PIR diverge from the predicted rates in PDB for disordered segments longer than 35. Disordered molecules rarely form crystals. Thus, this divergence of PDB from SW and PIR probably results from the inhibition of protein crystallization when LDRs are present, thus leading to a bias in PDB against proteins with LDRs.

The frequencies of predicted DRs in SW and PIR are well above estimates of false positives from the NRL_3D curves for all length classes, but especially in the LDR region. For example, for predicted DRs of 80 or longer, the LDR-based NN predictor gave 0.3% false positives error rates on NRL_3D (rounded to 0% in Table 2), whereas it gave 7% on SW and 6% on PIR. Application of the rule-based predictor to the same problem gave 3% false positives and 17% for SW and 14% for PIR. Perhaps our current implementation of the LDR neural network predictor is overly stringent and misses substantial numbers of LDRs. Overall,

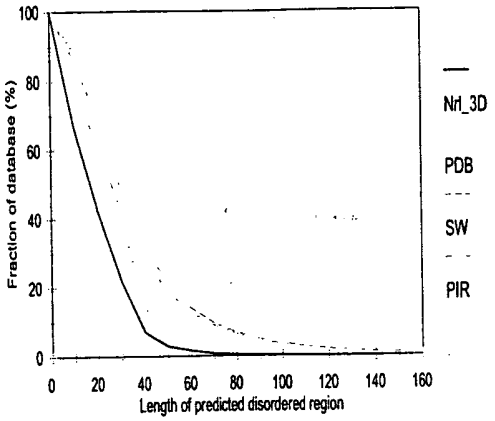

Figure 2: Comparison of LDR-based NN prediction on 4 major databases

Predicted DR longer than:	Fraction of proteins over whole database			
	NRL_3D	PDB	SW	PIR
20	42%	65%	70%	62%
40	7%	14%	32%	28%
60	2%	4%	14%	13%
80	0%	1%	7%	6%
100	0%	1%	4%	3%
120	0%	1%	2%	2%
140	0%	0%	1%	1%
160	0%	0%	1%	1%
180	0%	0%	0%	0%

Table 3: Fraction of proteins predicted to have DR longer than specific values.

these observations suggest that DRs, even LDRs, are very common in nature.

5 Conclusions

The LDR, MDR and SDR predictors were significantly more accurate when applied to the same length class. Even the process of grouping all disordered amino acids together reduced the prediction accuracy substantially (Table 2). These results strongly suggest that amino acid sequence characteristics leading to disorder are dependent on the length of the disordered region.

The current views of protein structure and function still seem to be dominated by the concepts of rigid organization and lock-and-key interactions [6], despite many examples of disorder-to-order transitions upon binding. As we point out, disorder-to-order transitions upon binding have been found for a diversity of molecular interactions that span the biological domain [4].

Koshland's induced-fit hypothesis introduced flexibility as an alternative to the lock and key [10]. Without reference to induced fit, Schulz [13] pointed out that the increases in free energy when flexible components solidify upon binding enable high specificity without excessively high affinities. Petsko and his collaborators [1] independently showed that loss of flexibility could help prevent excessively tight binding, but failed to note the coequally important feature of trading flexibility for specificity. We recently extended Schulz's proposal to show that disorder-to-order transitions allow natural selection to operate separately on affinity and specificity. We propose that such a separation is essential for the evolution of complex signaling and metabolic networks [4].

If disordered regions are required for the separation of affinity and specificity as we propose, then such regions should be very common. The commonness of such regions is fully supported by the findings presented herein. The next steps will be to determine whether the regions predicted to be disordered do indeed carry binding function and to determine whether the predictions are correct. Studies in these directions are underway.

Here we assume that all invisible regions in the x-ray structures are equivalent; however, three different causes have been identified for such regions, including crystal packing disorder, static disorder, and dynamic disorder [8]. Only the last of these involves the local disorder required by Schulz's proposal, so lumping all invisible regions together as we have done may be inappropriate. This is an acknowledeged weakness of the present study and might be adding noise to the data. We plan to improve our labeled data sets by distinguishing the 3 types of disorder for as many entries as possible, using either literature-based investigations or laboratory-based experiments.

Due to the curse of dimensionality and the small sizes of the DR labeled data sets, our current neural network predictors were very simple and limited to just a few features. Yet fairly good predictions are evidently being made despite this limitation. Increasing the sizes of DR labeled data sets and repeating the feature selection process for each length class will enable us to test a larger pool of candidate features and to design more complex predictors.

The more complex predictors will, hopefully, give more accurate predictions. This is especially important for short disordered segments since none of the current predictors do very well in this region. Since SDRs are found frequently to be involved in disorder-to-order transitions upon binding [4], improved predictions in this domain are certainly important.

Acknowledgments

Susan Johns and Steve Thompson of the Center for Visualization, Analysis and Design in the Molecular Sciences at WSU are acknowledged for their help with the various molecular biology data bases and with the use of the GCG package for calculating some sequence features. This work was sponsored in part by a grant from the American Heart Association, Washington State Chapter awarded to A. K. Dunker and Chul-Hee Kang. Molecular Kinetics is thanked for providing additional support.

References

[1] Alber, T. et al. [1982] "The role of mobility in the substrate binding and catalytic machinery of enzymes," *Mobility and function in proteins and nucleic acids,* Pitman, London, pp. 4-24.

[2] Anfinsen, C. B. [1973] "Principles that govern the folding of protein chains," *Science* vol. 181, pp. 223-230.

[3] Burley, S. K. and Petsko, G. A. [1985] "Aromatic-aromatic interaction: A mechanism of protein structure stabilization," *Science,* vol. 229, pp. 23-28.

[4] Dunker, A. K., Romero, P., Obradovic, Z., Kissinger, C. R., and Villafranca, J. E. [in preparation] "Role of protein disorder in the evolution of molecular recognition."

[5] Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. [1982] "The helical hydrophobic moment: a measure of the amphiphilicity of a helix," *Nature,* vol. 299, pp. 371-374.

[6] Fischer, E. [1894] "Einfluss der configuration auf die wirkung derenzyme," *Ber. Dt. Chem. Ges.* vol. 27, pp. 2985-2993.

[7] Fukunaga, K. [1990] *Introduction to statistical pattern recognition,* Academic Press, San Diego, CA.

[8] Huber, R. [1979] "Conformational flexibility and its functional significance in some protein molecules," *TIBS,* vol. 4, pp. 271-276.

[9] Kissinger, C. R. et al. [1995] "Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex," *Nature,* vol. 378, pp. 641-644.

[10] Koshland, D. E. [1958] "Application of a theory of enzyme specificity to protein synthesis," *Proc. Nat'l. Acad. Sci. USA* vol. 44, pp. 98-104.

[11] Kyte, J. and Doolittle, R. F. [1982] "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol,* vol. 157, pp. 105-132.

[12] Muchmore, S. W. et al. [1996] "X-ray and NMR structure of human Bcl-xl, an inhibitor of programmed cell death," *Nature,* vol. 381, pp. 335-341.

[13] Schulz, G. E. [1979] "Nucleotide binding proteins," *Molecular Mechanism of Biological Recognition,* Elsevier/North-Holland Biomedical Press, pp. 79-94.

[14] Vihinen, M., Torkkila, E. and Riikonen, P. [1994] "Accuracy of Protein Flexibility Predictions," *Proteins: Structure, Function, and Genetics,* vol. 19, 1994, pp. 141-149.

[15] Werbos, P. [1974] "Beyond regression: New tools for predicting and analysis in the behavioral sciences," Harvard University, Ph.D. Thesis. Reprinted by Wiley and Sons, 1995.

Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization

Ethan Garner [1] egarner@wsunix.wsu.edu
Paul Cannon [2] pcannon@eecs.wsu.edu
Pedro Romero [2] promero@eecs.wsu.edu
Zoran Obradovic [2] zoran@eecs.wsu.edu
A. Keith Dunker [3] dunker@mail.wsu.edu

[1] Department of Biochemistry and Biophysics, Washington State University, Pullman, WA 99164-4660
[2] School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-4660
[3] Author to which all correspondence should be addressed.
Department of Biochemistry and Biophysics, Washington State University, Pullman, WA 99164-4660
Telephone: 509 335-5322, Fax: 509 335-9688

Abstract

Using ordered and disordered regions identified either by X-ray crystallography or by NMR spectroscopy, we trained neural networks to predict order and disorder from amino acid sequence. Although the NMR-based predictor initially appeared to be much better than the one based on the X-ray data, both predictors yielded similar overall accuracies when tested on each other's training sets, and indicated similar regions of disorder upon each sequence. The predictors trained with X-ray data showed similar results for a 5-cross validation experiment and for the out-of-sample predictions on the NMR characterized data. In contrast, the predictor trained with NMR data gave substantially worse accuracies on the out-of-sample X-ray data as compared to the accuracies displayed by the 5-cross validation during the network training. Overall, the results from the two predictors suggest that disordered regions comprise a sequence-dependant category distinct from that of ordered protein structure.

1 Introduction

Many regions of proteins and some whole proteins form ensembles of structures under native conditions, in essence lacking a fixed tertiary structure within a given functional domain. Such "disordered" (or "unfolded") proteins have been identified by several methods: 1. sensitivity to proteases; 2. missing electron density in structure determinations by X-ray diffraction; 3. NMR spectroscopy; and 4. CD spectroscopy coupled with other methods such as rapid protease digestion, gel exclusion chromatography, or survival of function following incubation at high temperatures.

Disordered regions characterized by the methods described above are often essential for function. Such regions have therefore been called 'natively unfolded' [51], or 'natively disordered' [20]. 'Unfolded' implies that the region of protein exists in an extended, flexible (random-coil-like) form, whereas 'disordered' includes not only these extended forms but in addition can also imply a collapsed, partially folded with secondary structure, but non-rigid (molten-globule-like) form. The disordered ensemble of structures can involve equilibria between random-coil-like and molten-globule-like forms.

Since amino acid sequence determines protein structure [2], we proposed that amino acid sequence should determine lack of tertiary structure or disorder as well [43]. To test this hypothesis, we identified disordered sequences from missing coordinates in Protein Data Bank (PDB) files and then developed

and tested a collection of neural network predictors of order and disorder [44, 43, 45]. Following our initial studies on disordered regions of various lengths, our focus turned to long disordered regions (LDRs), where long is somewhat arbitrarily defined as > 40 contiguous amino [45]. Such long disordered regions have lower false positive error rates and are unlikely to be completely missed, even for a modestly accurate predictor. The first generation neural network predictor (NNP) of such long regions, called the LDR NNP but herein renamed the X-RAY NNP, was shown to predict order and disorder on a residue-by-residue basis with 73% accuracy as estimated by 5-cross validation. The false positive error rate for contiguous predictions of 40 or longer was found to be less than 7% on a sequence basis [43], which corresponds to less than 1 false positive prediction of disorder equal to or longer than 40 for every 1,000 amino acids [45].

Missing electron density in an X-ray-solved protein structure can result from experimental limitations or can be the result of structural disorder. Structural disorder can be static or dynamic [7, 25], which considerably weakens confidence in predictions based on this data. Thus, it would be useful to obtain data from other methods or at least to use X-ray data verified by other methods.

Several proteins with disordered regions have been characterized by NMR spectroscopy. Since this method does not suffer from an uncertainty regarding the type of disorder, application of our X-RAY NNP to several LDRs characterized by NMR provides a means to stringently test the validity of our predictor. Also, further validation of the X-RAY NNP training proteins could be accomplished by testing them with a predictor trained on NMR-characterized proteins, so a second predictor, herein called the NMR NNP, was developed. The two neural networks were then tested upon each other's training sets. Although the accuracies of the two predictors upon each other's set of disordered regions showed large variations, further study suggests that the variable accuracies are the result of different types of disordered regions. Overall, the results presented herein suggest that disordered or unfolded regions of sequence form a distinct category compared to ordered or folded regions of sequence.

2 Materials and Methods

2.1 The Proteins

In our previous work [43] disordered regions were identified from Protein Data Bank (PDB) [1] files as amino acids that were missing from the set of atomic coordinates. That is, disorder leads to incoherent X-ray scattering and subsequent absence of electron density in the solved structure [7, 30, 47]. These proteins containing X-ray-characterized regions of disorder, their PDB filenames, and their Swiss Protein [6] identifications (SW IDs), respectively were; 1. Tomato Busy Stunt Virus, 2tbv, COAT_TBSVB, 2. Tyrosyl tRNA synthetase, 2ts1, SYY_BACST, 3. Calcineurin, 1aui, P2BA_HUMAN, 4. Topoisomerase II, 1bgw, TOP2_YEAST, 5. Elongation factor G, 1elo, EFG_THETH, 6. Apoptosis regulator BCL-Xl, 1bcl, BCPA_PROAE, 7. Intact lactose operon repressor, 1lbh, LACI_ECOLI. The total number of disordered amino acids in this database is 449 aa.

NMR proteins with identified regions of disorder were identified by tedious literature searches. Several different NMR parameters identify regions of disorder [41, 54, 19, 55]. The proteins and their SW (or PIR) IDs were 1. 4e binding protein 1, s50866 (PIR), 2. Murine Prion, PRIO_MOUSE, 3. Histone H5, H5_CHICK, 4. Flagellum specific sigma factor (FlgM), FLGM_SALTY 5. Antitermination protein of bacteriophage λ (AT), REGN_LAMBD, 6. N term activator domain of Heat Shock Transcription Factor (HSTF), HSF_KLULA, 7.High Mobility Group-I (HMG-I), HMGI_HUMAN. The total number of disordered regions in this database is 677 aa.

In addition, we collected a similar number of structured control proteins. These proteins were selected to be of similar overall size as the X-ray and NMR proteins, to be monomeric, and to be without cofactors. These proteins, their PDB filenames and their SW Ids were, respectively; 1. Hen egg-white lysozyme; 1hel, LYC_CHICK, 2. Ribonuclease A (Rnase A), 3rn3, RNP_BOVIN, 3. β-cryptogein (B-cryp), 2ctb, CBPA_BOVIN, 4. Elastase, 1lvy, EL1_PIG, 5. Profilin A (Pfln A), 1acf,

PRO1_ACACA, 6. Haloalkane Dehalogenase (HDHase), 2edc, HALO_XANAU, 7. Azurin II (Az II), 1arn, AZU2_ALCXX, and 8. Carboxypepitidase A (CbPA), 2ctb, ELIB_PHYCR.

2.2 Feature Selection

We use the term 'attribute' to mean a value calculated over a specifies window and the term 'feature' for those attributes that are subsequently used to train the neural networks. Sequence attributes are numerical values calculated from an amino acid sequence over a specified window [4]. For these studies, 24 attributes provided the initial pool, the first 20 of which are the compositions of the 20 amino acids within the specified sequence windows. The last 4 are hydropathy [33], flexibility index [50], helix amphipathic moment [21] and sheet amphipathic moment [22].

The NMR and X-ray disordered datasets were each matched with an equal number of ordered amino acids taken from the NRL_3 [40], which is a subset of PDB containing only ordered structures. A feed-forward search with minimal error probability selection criterion was used on the balanced ordered and disordered NMR and X-ray datasets [43]. A quadratic Gaussian classifier using different covariance matrices for each class was used to calculate the minimal error probability during each of the searches. Experimentation with other dimensionality reduction methods, such as sequential backward search and branch-and-bound, yielded results quite similar to those presented here. Ten features were selected from the original pool of 24 attributes.

2.3 Neural Network Training

Several possible neural network architectures were investigated in the initial phase of these studies. A simple network with 10 inputs, 7 fully connected nodes in a single hidden layer, and one output was selected as being commensurate with the dataset size and as giving good results [43].

The X-ray and NMR disordered datasets, with their number-balanced datasets of ordered sequences, were scrambled in order to separate values from adjacent sequence positions, and then divided into 5 disjoint subsets by random selection. Experimentation indicated that similar prediction accuracies were achieved during training whether or not scrambling was used, but scrambling may serve to improve predictions for completely unrelated proteins.

For each training cycle, 4/5 of the data comprised the training set and 1/5 the test set. The training set was further separated into a proper training set (80%) and a validation set (20%). Three initializations were used and the number of epochs for each training was chosen as that which produced the highest accuracy on the validation set. Once training was investigated by 5-cross validation, the data were recombined and training was repeated using 5/5 of the data.

3 Results

3.1 Selected Features

The ten features selected on the basis of distinguishing order and disorder for the NMR and X-ray datasets are shown in Table 1. Six of the ten features were the same for both datasets: flexibility index, hydropathy, and mole fractions of Y, W, C, and S. These data indicate that the NMR-and-X-ray-characterized regions of disorder share important characteristics.

With regard to the selected features that were different for the two datasets, the compositions of H, D, K, and E distinguished ordered and disorder for the X-ray dataset, whereas the compositions of F, G, R, and P were useful for the NMR dataset. Thus, the features selected for the NMR-characterized regions of disorder show important differences from those selected for the X-ray-characterized regions of disorder.

X-RAY	H	D	K	E	S	C	W	Y	Hydropathy	Flexibility
NMR	F	G	R	P	S	C	W	Y	Hydropathy	Flexibility

Table 1: **Selected features.**

X-RAY NNP	NMR NNP
70%	86%
77%	84%
77%	89%
72%	86%
70%	89%
Average 73% ± 2%	Average 87% ± 4%

Table 2: **Five Cross Validation Results.**

3.2 Five Cross Validation

The evaluation of the training of the X-RAY NNP was described previously [43]. Here those data are compared with the results of a similar training exercise for the NMR NNP (Table 2.). Overall, the NMR NNP gives a significantly higher accuracy compared to the X-RAY NNP during the training exercises, e.g. 87%± 4% compared to 73%±2%.

3.3 Example Predictions

Example predictions are shown in Fig. 1. The X axis is the residue number while the Y axis is the prediction output. Anything above an output of 0.5 is considered a prediction of disorder. The solid horizontal line at the center of the graph indicates what regions are actually disordered. Fig. 1A and Fi. 1B are predictions on disordered proteins, while 1C and 1D are predictions upon the ordered control proteins. One of the best overall predictions (1A) and one of the worst (1B) on regions of disorder as well as the two worst overall predictions (1C, 1D) on the control proteins are provided. In (1A), the prion protein from the NMR dataset was subjected to analysis using both the NMR NNP and the X-RAY NNP. Notice how the X-ray prediction accuracy is relatively similar to that of the NMR predictor, which was trained on this protein's data (X-RAY NNP = 88.4% correct overall; NMR NNP = 97% correct overall). In (1B), the anti-termination (AT) protein from bacteriophage lambda from the NMR dataset was predicted upon by both predictors. Here, the accuracy of the X-RAY NNP (53.2%) seems very poor, but notice how its prediction is again somewhat similar to that of the NMR NNP, which, despite having this protein in its training set, still manages an accuracy of only 73%, much lower than that obtained on the prion protein in the previous example. Finally, in (1C) and (1D) the predictions on the ordered control proteins, profilin A and haloalkane dehalogenase, are presented, respectively. The false positive predictions of disorder are seen to be very short, especially for the NMR NNP.

For long disordered regions (LDRs) such as that for the AT protein, even modestly successful prediction rates (e.g. just 53.2% for the X-RAY NNP) still give an indication of protein disorder. For this reason, we are initially focussing our attention on proteins with such LDRs.

Fig. 1 also indicates several types of errors. Relative to prediction of disorder, false positive predictions are ordered regions incorrectly predicted to be disordered (peak labeled b in 1A) whereas false negative predictions are disordered regions predicted to be ordered (region a in 1A, regions c, d, and e in 1B and so on for 1C and 1D). Another useful classification is whether an errant prediction is false (for example peak a in 1A) throughout (e.g. a non-boundary error, for example peak b in 1A) or is correct over some region but then becomes false upon crossing an order/disorder junction in the

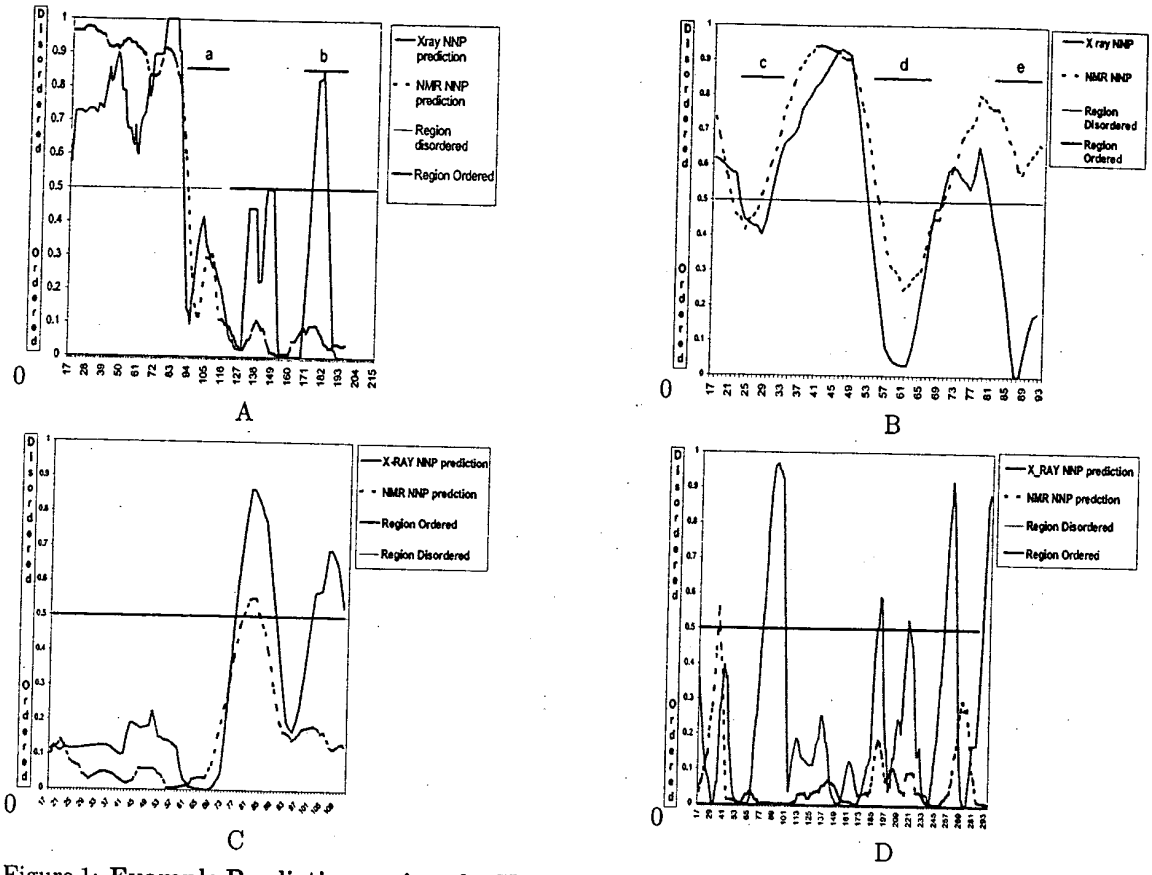

Figure 1: **Example Predictions using the X-RAY NNP and NMR NNP.**– Example predictions using the X-RAY and NMR NNPs. Both predictors were applied to the following proteins: murine prion (A); bacteriophage λ antitermination protein (B), profilin A (C) and haloalkane dehalogenase (D). The first two contain regions of disorder and the last two are ordered, control proteins. The X-axis is the residue number, while the Y axis is the prediction output. Values above 0.5 indicate disorder, below 0.5 indicate order. The solid line at 0.5 indicates an identified region of order, a dashed line a region of disorder. Various types of errors are marked and indicated by letters a, b, c, etc. (see text).

structure (e.g. a boundary error).

3.4 Prediction Accuracies

The X-RAY NNP was applied to the NMR-characterized proteins and the NMR NNP was applied to the X-ray-characterized proteins. The results of these out-of-sample predictions are presented in Table 3.

For the X-RAY NNP, the overall prediction accuracies range from 53.2% (AT) to 93.5% (HMGI(Y). The large range of error rates undoubtedly relates to a variation in the degree of similarity of the disordered regions in the different proteins to the disordered regions used to train the X-RAY NNP. For example, unlike most of the NMR proteins, the HMGI(Y) has local charge imbalance, thus having charge attributes commensurate with those of the X-ray training set and giving a very high overall prediction accuracy by the X-RAY NNP.

For the NMR NNP, the overall prediction accuracies range from 55.1% (tyrosyl-tRNA synthetase) to 94.1% (Bcl-xL). Again, the low prediction accuracy on the synthetase signals a difference in the

X-RAY NNP on NMR-Characterized Data

Protein	Length of sequence	Prediction of Disordered Regions (DR) from amino acid sequence: Known disorder	Region predicted	Predicted DR lengths	Percent correct	False negative	False positive	Structural characterization	Ref.
Antitermination protein of bacteriophage λ (ATa)	1-107	1-107	15-93	8, 23, 11	53.2%	46.8%	N/A	A, B, D	[36]
Histone H5 (H5)	1-189	1-21, 101-185	15-175	17, 16, 99	68.9%	0%	63.3%	A, C, D	[5]
Flagellum specific sigma factor (FlgM)	1-97	1-97	15-83	24, 37	88.4%	11.6%	N/A	A, D	[16]
N term activator domain of Heat Shock Transcription Factor (NAD-HSTF)	1-195	1-195	15-195	3, 58, 14, 31, 9	63.5%	36.5%	N/A	A, D	[14]
High Mobility Group-I (HMG-I (Y))	1-106	1-106	15-92	73	93.5%	6.4%	N/A	A, C, D	[31]
4e Binding Protein 1 (4e BP-1)	1-118	1-118	15-104	2, 15, 37	57.8%	42.2%	N/A	A, C	[26]
Murine Prion	1-254	23-120	15-240	73, 10	88.4%	22.5%	7.8%	A, D	[41]

NMR NNP on X-RAY-Characterized Data

Protein	Length of sequence	Prediction of Disordered Regions (DR) from amino acid sequence: Known disorder	Region predicted	Predicted DR lengths	Percent correct	False negative	False positive	Structural characterization	Ref.
Tomato Busy Stunt Virus	1-387	1-66	15-373	65, 17	73.9%	50%	17.3%	B, D	[29]
Ligase (tyrosyl tRNA synthetase)	1-206	109-206	15-192	18, 21	55.1%	74.6%	18.9%	B	[11]
Calcineurin	1-179	74-168	15-165	24	55.6%	73.6%	0%	B	[32]
Topoisomerase II	1-400	225-273	15-385	9, 4, 8	85.8%	83%	4%	B, D	[8]
Elongation factor G	1-341	50-125	14-327	39, 20, 6, 14, 14	75.1%	36.8%	21.1%	B	[23]
Apoptosis regulator BCL-X	1-196	1, 31-80	15-182	50	94.1%	11.5%	3.4%	B, D, A	[46]
Intact lactose operon repressor	1-360	1-61	15-346	11, 10, 17, 39	69.3%	76.5	23.1	B, D, A	[34]

Structural Characterization
A= NMR, B=X-ray diffraction, C= CD, D= Protease hypersensitivity.

Table 3: **Cross Prediction Results.**

characteristics of its disordered regions compared to those in the NMR dataset. The details of this difference await further study. On the other hand, the disordered region in Bcl-xL must be more similar to those in the NMR training set. It is likely to be coincidental that the structure of Bcl-xL has also been determined by NMR [37].

Application of the NMR and X-RAY NNPs to the control proteins was carried out. The results on a protein-by-protein basis are shown in Table 4. The error rate ranges from a low of 72.7% (X-RAY NNP Haloalkane Dehalogenase) on to a high of 100% (many proteins).

Finally, the overall prediction accuracies are summarized in Table 5. The NMR and X-RAY NNPs give similar overall prediction rates near 74% on each other's training sets. The predictions on the fully ordered control proteins are considerably better, about 84% for X-RAY NNP and 98% for the NMR NNP. The high accuracy on the fully ordered proteins implies that the ordered part of our training sets is providing our predictors with information that allows for better generalization than that achievable from our disordered data.

4 Discussion

4.1 X-ray- and NMR-Characterized Regions of Protein Disorder

Our pilot studies indicated a definite relationship between amino sequence and the presence of ordered or disordered structure [43, 44, 45, 20] However, these initial studies had two, interrelated, acknowledged limitations related to their exploratory nature. First, the number of disordered examples was very small, just 449 amino acids in the original LDR set. Second, all the disordered examples were from PDB, which has considerable ambiguity with regard to the characterization of disordered proteins. Fortunately, the signals for the tendency for disorder seem to be so strong that these two limitations apparently haven't led to large errors. Work in progress with a 6-fold larger database, now about 2,500 disordered amino acids, yield results very similar to those of the pilot studies (manuscript in preparation).

The small number of examples of disordered proteins in our original studies resulted from the

X-RAY NNP on Control Proteins

Protein	Length of sequence	Prediction of Disordered Regions (DR) from amino acid sequence: Known disorder	Region predicted	Predicted DR lengths	Percent correct	False negative	False positive	Structural characterization	Ref.
Hen egg-white lysozyme	1-129	None	15-115	4	96%	N/A	4%	B	[53]
Ribonuclease A (Rnase A)	1-124	None	15-110	3, 3	93.8%	N/A	6.2%	B	[10]
β-cryptogein (B-cryp)	1-98	None	15-84	None	100%	N/A	0%	B	[9]
Elastase	1-240	None	15-226	14, 7,4	87.4%	N/A	12.6%	B	[35]
Profilin A (Pfln A)	1-125	None	15-111	12, 11	75.3%	N/A	24.7%	B	[24]
Haloalkane Dehalogenase (HDHase)	1-310	None	15-297	22, 6, 2, 13, 7	72.7%	N/A	17.3%	B	[27]
Azurin II (Az II)	1-129	None	15-115	7, 10	83.2%	N/A	16.8%	B	[18]
Carboxypepitidase A (CbPA)	1-307	None	15-293	3, 9, 6, 9, 3, 2, 6, 21, 16	73.2%	N/A	26.8%	B	(not pub.)

NMR NNP on Control Proteins

Protein	Length of sequence	Known disorder	Region predicted	Predicted DR lengths	Percent correct	False negative	False positive	Structural characterization
Hen egg-white lysozyme	1-129	None	15-115	None	100%	N/A	0%	B
Ribonuclease A (Rnase A)	1-124	None	15-110	6	93.7%	N/A	6.3%	B
β-cryptogein (B-cryp)	1-98	None	15-84	None	100%	N/A	0%	B
Elastase	1-240	None	15-226	6, 1, 3	96.4%	N/A	3.6%	B
Profilin A (Pfln A)	1-125	None	15-111	6	93.8%	N/A	6.2%	B
Haloalkane Dehalogenase (HDHase)	1-310	None	15-297	1	99.6%	N/A	0.4%	B
Azurin II (Az II)	1-129	None	15-115	None	100%	N/A	0%	B
Carboxypepitidase A (CbPA)	1-307	None	15-293	0	100%	N/A	0%	B

Structural Characterization
A= NMR, B=X-ray diffraction, C= CD, D= Protease hypersensitivity.

Table 4: **Prediction on Controls.**

Overall Accuracy

Set	Total aa predicted on	Total ordered aa predicted on	Total disordered aa predicted on	Total false negative aa	Total false positive aa	Percent false negative	Percent false positive	Percent overall correct
X-RAY on NMR	884	207	677	179	60	26.4%	28.9%	72.9%
NMR on X-RAY	1873	1424	449	204	265	59.0%	14.3%	75.0%
X-RAY on Control	1238	1238	0	N/A	201	N/A	16.2%	83.8%
NMR on Control	1238	1238	0	N/A	29	N/A	2.4%	97.6%

Table 5: **Summary Tables.**

lack of organized data on non-folding amino acid sequences. PDB is the largest organized source of information about proteins with regions of disorder, but as our initial studies clearly demonstrate, PDB is strongly biased against the presence of disorder [44], so even this source does not have very many examples.

In addition to being few in number, X-ray-characterized regions of structural disorder have alternate possible causes for the observed missing electron density, including the following possibilities; 1. A locally *structured* domain could be moving; 2. a locally *structured* domain could be occupying several alternative positions; 3. a local region of sequence could comprise an ensemble of interconverting shapes; and 4. A local region of sequence could comprise an ensemble of static shapes. From our perspective, the important distinction is whether a region of sequence folds into a single structure (e.g. either 1 or 2) or comprises an ensemble of structures (e.g. 3 or 4). The distinctions between 1 and 2 and the distinctions between 3 and 4 are less important; indeed, 1 versus 2 and the 3 versus 4 distinctions are a continuum that depends only on the timescale. Here, we refer to locally structured domains that are disordered by movement or by occupancy of different positions (e.g. either 1 or 2) as 'domain wobble.'

Others have used 'dynamic,' 'static,' 'hinged' and 'flexible' to describe the various possible causes of structural disorder that leads to missing electron density in X-ray crystal structures [7, 25] but these previous terms do not correspond in any precise way to the 4 possibilities listed above. It is for this reason that we are proposing the terms 'domain wobble' and 'ensembles of structures' to contrast the distinctions that we believe are important to this work. In our initial studies we attempted to eliminate wobbly-domains by literature studies on each protein. However, not only does PDB lack clear information regarding disorder, but attempts to find information from the original literature often fail because the needed experiments have simply not been done or have not been reported. We hope that, as the importance of disordered protein becomes more generally realized, the key information

about such disordered regions will become more readily available.

Given the above, extending the studies of disorder to include regions characterized by NMR is important for overcoming both limitations of our initial studies. NMR-characterized disordered regions include information about the extent of folding of the disordered region and at the same time increase the number of examples.

Unfortunately, there are relatively few examples of NMR-characterized regions of disorder, and these are scattered in the literature and not collected at one location. So, just as for development of a disordered regions database using PDB, an intensive effort is required for each new entry characterized by NMR. The amount of effort required will continue to slow the rate of enlargement of our database of disordered protein. Nevertheless, the amount of disordered data in this paper has more than doubled the data compared to that of the pilot studies.

4.2 Feature Selection

Our initial work [45, 44, 43] emphasized the use of sequence attributes based on amino acid composition. We reasoned that LDRs could be considered to be a new "structural class," and amino acid compositions had been shown to be successful for protein class prediction [38]. We are aware that considerations of coupling effects among different amino acids has led to much improved prediction of protein class [15], and we would like to apply such approaches to disorder prediction. However, consideration of amino acid pair frequencies requires much more data than we currently have, so such approaches are simply out of reach at the present time.

The feature selection experiments in the development of the X-RAY and NMR NNPs (Table 1.) suggest substantial similarities for the disordered regions characterized by these two methods. While 6 out of 10 of the features are identical to each other, the remaining 4 contribute enough information to cause the differences noticed between the predictors. The fact that the NMR NNP has both a higher false negative rate and a lower false positive rate than the X-RAY NNP suggests that the NMR NNP has a higher threshold to which it ascribes its disordered features. This may be due to the fact that the NMR NNP's training set contains more extreme values for the attributes specific to disorder within its training set (see Fig. 1), values not found as frequently within the X-ray data set correlating with disorder/order predictions .

We have developed a substantially larger database, having approximately 2,500 disordered amino acids in windows of 21 matched with an equal number of ordered amino acids in windows of the same size. Studies in progress on this larger database indicate that charge imbalance, when it exists, is a very strong determinant of local disorder (manuscript in preparation). The selection of H, D, K and E for the X-ray dataset is the result of substantial charge imbalance in several of the disordered regions in the X-ray-characterized proteins. In contrast, charge imbalance is not so important for the current NMR dataset.

Flexibility index, hydropathy and the mole fraction of S were found to be relatively higher in disordered regions as compared to ordered regions for both the NMR and X-ray data in the enlarged dataset, which is in complete agreement with the pilot studies [44, 42, 43, 45]. More flexible, more polar regions are more likely to be disordered. Not only does S promote disorder by its polarity, but it is special owing to its generally high abundance coupled with its ability to stabilize multiple local backbone conformations by side-chain-backbone hydrogen bonding [48].

On the other hand, Y, W, and C were lower in the disordered regions as compared to the ordered regions, both in the new enlarged database and in the data used for pilot studies. In several different datasets of ordered and disordered regions, W, Y, and C have always been found to be lower in disordered regions: indeed, these three appear to be the most order-forming of the natural amino acids (manuscript in preparation). The order-forming tendencies of W and Y may be related to the extra stability arising from aromatic/aromatic interactions [12], while the ability to form disulfide bonds is an obvious reason for the order-forming potential of C. Interestingly, W, Y, and C also

evidently have the highest tendency to be conserved as judged, for example, by the various values in the PAM 250 matrix [17].

With regard to the features selected for the NMR dataset, the studies in progress on the larger dataset (manuscript in preparation) suggest that disorder is found to be associated with high mole fractions of R and P and with low fractions for F. R is typically an uncommon amino acid, but when there are high local concentrations R, it likely induces disorder by charge imbalance. High levels of proline prevent compact folding; indeed, proline-rich regions are common in proteins and seem to have function associated with their ill-folded conformations [52, 3, 39]. Higher local concentrations of F probably encourage order for the same reasons as Y and W - due to extra stability from aromatic/aromatic interactions [12].

In the study on the larger dataset mentioned above, the mole fraction of G when considered alone is found to be essentially uncorrelated with either order or disorder, so it is unclear why this amino acid was selected for the NMR dataset. On the other hand, in the development of flexibility index, G was found to change markedly, showing high flexibility index values when next to flexible neighbors, but showing low flexibility values when next to less flexible neighbors [50]. Thus, the selection of G may reside in its behavior in conjection with neighboring amino acids in the ordered and disordered sequences. This is consistent with the way feature selection works, trying to select not only the best features but also the best combinations of features. The mole fraction of G may not be correlated with order/disorder, but its combination with other variables improves the discriminatory power of the predictor.

Since the NMR NNP has a lower rate of false positives than the X-RAY NNP we plan to be use it to complement other NNP's by helping weed out false positive predictions. Future predictors based upon a larger training set of diverse proteins may yield better results, as characteristics of disorder from differing families of proteins are incorporated into the predictors. A new predictor is currently under development that combines both of these training sets, and selects from a greater number of features (Unpublished). Preliminary results show greater accuracy as judged by 5-cross validation, suggesting that enlarging the data set can lead to greater prediction precision as more features indicative of disorder are included. Also, a larger data set can provide our predictors with enough disorder data to allow for better generalization.

4.3 Out-of-Sample Predictions

The X-RAY NNP and the NMR NNP give similar overall prediction accuracies on the proteins used in each other's training sets: 72.9% for the X-RAY NNP on the NMR data and 75% for the NMR NNP on the X-ray data. However, significant differences become evident when the types of errors are considered.

The X-RAY NNP on the NMR data exhibits similar false positive (28.9%) and false negative (26.4%) rates, whereas the NMR NNP on the X-ray data exhibits large false negative rates (59%) and small (14.3%) false positive rates. Additional insight follows from noting that the more or less uniform performance of the X-RAY NNP on the NMR data with a overall accuracy of 72.9% closely matches the 5-cross validation results (73%), whereas the much more variable performance of the NMR NNP is associated with a large drop-off in the out-of-sample predictions (about 75%) as compared to the 5-cross validation results (87%). Overall, these data suggest that the NMR NNP is much more specific for the disordered regions characterized by NMR, whereas the X-RAY NNP appears to be more general.

One possibility for the poor performance of the NMR NNP on the X-ray-characterized disordered regions that these disordered regions are misclassified, e.g. they are actually wobbly domains.

For example, a region of missing electron density in a tyrosyl t-RNA synthetase different from the one in the present studies was later shown to have a considerable part that is ordered [28]. This observation on another t-RNA synthetase coupled with the high false negative error rate (74.6%) for the

NMR NNP (Table 3.) could be an indication that we have misclassified the disordered region identified by X-ray diffraction in this protein. On the other hand, the NMR NNP's worse false negative prediction (e.g., 83% for topoisomerase II) would seem to indicate that this disordered region surely must be misclassified. However, the disordered region of this protein has been well-characterized to lack ordered structure: the putatively disordered region is extremely rich in charged residues and hypersensitive to protease digestion [13, 49]. Indeed, most of the disordered regions of the X-ray characterized data exhibit hypersensitivity to protease digestion at multiple sites, which argues strongly against ordered structure for these regions. The NMR NNP shows very poor performance on several of the X-ray-characterized regions of disorder with false negative values well over 50%. The overall accuracies of the NMR NNP on these proteins gives reasonable values, above 50% in every case, because the predictions on the ordered parts of these proteins have good accuracies.

An alternative possibility to explain the very poor predictions of the NMR NNP on the X-ray-characterized proteins is that their disorder depends on charged residues, which are utilized by the X-RAY NNP but not by the current NMR NNP. As mentioned above, the disordered region of topisomerase II is highly charged and therefore consistent with this suggestion. Examination of the other disordered regions with high false negative error rates shows that these regions are also highly charged.

4.4 Control Predictions

The false positive rates of 16.2% (X-RAY NNP) is much lower than the evaluation of this same predictor on NRL_3D which gave a false positive error rate of 31.5%. The reasons for this large discrepancy are unclear, but may relate to the selection criteria for the control proteins in this study (e.g. small overall size to match the sizes, monomeric, no ligands, except for the metal ion in carboxypeptidase). NRL_3D contains large proteins, oligomeric proteins, and proteins with bound co-factors. All of these factors could contribute to false positive predictions. For example, a local tendency for disorder would be more likely to be over-ridden by non-local interactions in a larger protein as compared to a smaller one simply by chance, so larger proteins are more likely to have false positive predictions of disorder. We are testing this possibility. Oligomeric proteins might associate via disordered regions, in which case a prediction of disorder would appear as a false positive in the crystal structure of the oligomer. Finally, binding of co-factors could involve disorder-to-order transitions, so again predictions of disorder would appear as false positives in the structure of the holoprotein.

The NMR NNP exhibited an especially low false positive error rate on the control proteins, just 2.4%. This is more than 6 times smaller than the false positive error rate, 14.3%, on the ordered parts of the X_RAY NNP training set. For the NMR NNP on the X-RAY NNP training set, most of the false positive errors related to improper placement of the order / disorder boundary. Because of our windowing procedure, disorder information is carried into the ordered regions with the resulting tendency that the NMR NNP predicts disorder farther along the sequence than it actually occurs when a region of disorder is present. A further possibility is that in solution the regions of disorder actually extend farther along their respective sequences than indicated in the X-ray structures, which are more ordered than the solution state due to the ordering effects of crystal formation.

4.5 Summary

An NMR NNP and an X-RAY NNP were developed and tested on each other's training set proteins. The X-RAY NNP seems gives similar results across the ordered and disordered regions for both its own training set (as measured by 5-cross validation) as for the out-of-sample NMR NNP training set. In contrast, the NMR predictor does much better on its own training set (as measured by 5-cross validation) as compared to the out-of-sample X-RAY NNP training set. These data support the validity of the X-RAY NNP as a general predictor of protein disorder, suggesting that the uncertain interpretation of disorder characterized by X-ray diffraction in principle did not lead to a significant problem. On the other hand, the NMR NNP appears to be much more specific, for reasons that are

Protein Structure Prediction and Design

Veronica Morea[a,b], Raphael Leplae[a] and Anna Tramontano[a, *]

[a]IRBM P. Angeletti, via Pontina Km. 30.600, 00040-Pomezia (Rome), Italy

[b]Istituto di Chimica del Farmaco, Universita' "G. D'Annunzio", Chieti, Italy

*To whom correspondence should be addressed

Abstract

Proteins have a unique native conformation, which can be proven in many instances to be determined by the amino acid sequence alone. The folding problem, that is the understanding of how the amino acid sequence directs folding, is still unsolved, despite more than 30 years of efforts. However, many new methods have appeared in the past few years. This chapter describes the different principles underlying them and tries to give an overview of their successes and pitfalls.

Abbreviations

1D	mono-dimensional
3D	three-dimensional
CASP	Critical Assessment of Structure Prediction
PDB	Protein Data Bank
r.m.s.	root-mean-square
r.m.s.d.	root-mean-square deviation
URL	Universal Resource Location

INTRODUCTION

The information contained in known protein structures can be of invaluable help both to understand the function of individual proteins, for example to explain on a chemical basis the catalytic activity of enzymes, and to infer the general principles determining protein folding. The knowledge of the 3D structure of proteins is essential to understand their functions and/or properties and to be able to modify them in a predictable way. There are a number of successful examples where specific properties of proteins have been modified, by designing novel proteins, novel ligands or peptidomimetics. Modified proteins have been used, for example, to mimic ("agonists") or hamper ("antagonists") the action of a given ligand at the receptor level [1-3]. In some cases novel sequences and therefore novel structures have been synthesised to achieve specific tasks, or to be used as an appropriate scaffold for a given function ("*de novo* design") [4-14]. Peptidomimetics [15, 16] have been shown to be able to reproduce or antagonise the action of a protein by mimicking the structural elements involved in the recognition process. Small organic molecules able to bind to a target protein and inhibit its activity are often designed on the basis of the 3D co-ordinates of the binding site of a given receptor or enzyme [17-19].

More than five thousands protein structures are publicly available as of today [20] and, due to the continuous progress in X-ray crystallography and NMR spectroscopy, their number is increasing more and more rapidly. However, the number of known protein sequences is at least one order of magnitude higher and the gap between the number of known structures and sequences is continuously increasing, as a consequence of improvements in the methods of sequence determination and of the many ongoing genome projects [21] (see

URL: http://www.sanger.ac.uk/Projects/). Consequently, the prediction of protein structure from their amino acidic sequence represents an appealing perspective, which has been intensively pursued in the last three decades.

Although it is known that the amino acid sequence of a protein (primary structure) contains sufficient information to determine its 3D or tertiary structure [22], the specific mechanisms underlying protein folding are still eluding our understanding [23] and a multitude of different methods are continuously developed to try and predict a protein structure starting from its sequence.

Some of these methods generate a high number of possible conformations for a given protein sequence and try to select the conformation corresponding to the lowest energy (*'ab initio'*). Other methods, based on the assumption that protein folding is a process under kinetic, rather than thermodynamic, control, try to simulate the folding pathways of a protein. Both types of methods are very general and, if successful, would allow to predict the native conformation for any given protein sequence. Unfortunately, they have not been very successful up to now because of the complexity of the problem: the number of possible conformations of an average protein is extremely high and it is very difficult to obtain an accurate representation of all the physical forces acting on proteins.

Different methods have therefore been developed to assess if a given sequence is likely to assume a structure similar to that of an already known protein. These methods can be ascribed to one of two large categories: 'modelling by homology' (or 'comparative modelling') and 'fold recognition'. They are less general than *ab initio* methods in that they require the unknown protein to be similar in structure to an already known protein and need an efficient way to recognise this similarity; however, their results are very satisfactory in many cases.

All these methods aim at providing a model of the 3D structure of the whole protein. When this results unfeasible, however, it is still possible to attempt a partial prediction of the protein structure, for example by identifying its secondary structure elements.

Behind the continuous increase in the power of computing tools and the development of always new methods for protein structure prediction, two events in the last few years determined a considerable step forward in this field:

?? The free diffusion through the Internet of most of the available data on protein sequences and structures (see URL: http://www.embl-heidelberg.de/srs/srsc) and of the methods for protein structure prediction has given a great advantage to the community of the 'predictors' and has also allowed non-experts in the field to use the available methods via appropriate servers.

?? The two protein structure prediction competitions which have been held in 1994 and 1996 have served as an objective test to evaluate most of the published methods, highlighting their strengths and weaknesses, and providing the basis for further improvements.

In this chapter we will summarise the current situation in protein structure prediction and some of the implications for protein design. This chapter is by no means intended to provide an exhaustive list of the available methods; we will try instead to describe the principles underlying them and to highlight their strengths and limitations. We will mostly limit our description to prediction methods tested in the two protein structure prediction experiments, as rigorous blind testing is the only unbiased way to evaluate their performance.

THE CRITICAL ASSESSMENT OF TECHNIQUES FOR PROTEIN STRUCTURE PREDICTION (CASP) EXPERIMENTS

In 1994 and 1996 two large-scale experiments to critically assess the state of art in protein structure prediction have taken place (see URL: `http://PredictionCenter.llnl.gov/`) . These experiments consisted into two phases. First, X-ray crystallographers and NMR spectroscopists were asked to provide information about structures which were about to be solved or which had already been solved but not yet publicly disclosed. Second, the scientific community was asked to submit predictions for one or more of the target proteins. These predictions were subsequently compared to the experimental structures.

The predictions were divided into three categories according to the method used:

1. Comparative modelling
2. Fold recognition
3. *Ab initio* predictions

(In the second experiment a 'docking' category was also present but it will not be discussed here).

The predictions were assessed by independent teams, one for each category and meeting were held in December 1994 and 1996 "to examine what went right with the predictions, what went wrong, and, where possible, to understand why" [24].

There are obvious limitations to the significance of the results: the targets certainly do not represent a statistically unbiased sample of all possible protein structures; some groups only submitted a small number of predictions, the time allowed for the prediction was limited, different groups put different effort in the experiment and different methods were at different stages of development. Nevertheless these experiments still provided an objective picture of the capabilities and the deficiencies of most of the existing methods. From this picture, it became evident in which fields an improvement was mostly needed and that algorithms claiming predictive capabilities should be asked to demonstrate them through blind testing.

MODELLING BY HOMOLOGY AND LOOP PREDICTION

I. RATIONAL BASIS OF THE METHOD: RELATIONSHIP BETWEEN SEQUENCE IDENTITY AND STRUCTURAL SIMILARITY

'Modelling by homology' or 'comparative modelling' consists, in very simple terms, of two steps: 1) identification of the protein(s) of known structure ('parent') whose sequence is most similar to that of the protein to be predicted ('target'); 2) building of a model of the structure of the target protein using that of the parent protein(s) as a 'template'.

The rationale for this procedure is that there is a clear relationship between sequence identity and structure similarity in proteins: it has been shown [25] that the similarity of the backbone conformation in the core regions of two proteins increases with the sequence identity between them. In particular:

?? For proteins with sequence identity = 50% the r.m.s.d. of the backbone atoms of the core region[1] is = 1.0Å and this region comprises about 90% of their structure.

---

[1] The core region is defined by superimposing the backbone atoms of the secondary structure elements, and extending these elements to include additional residues at their ends, as long as the r.m.s.d. is = 3.0 Å [25]. The percentage of residues in this region depends upon the structural similarity between two proteins and, consequently, upon their sequence similarity.

?? For proteins with sequence identity = 20% the core region could comprise only 50% of their structure with an r.m.s.d. of the backbone atoms in this region > 1.8Å; relevant structural differences can occur outside of the core.

?? Proteins with sequence identity between 20% and 50% have an intermediate degree of similarity between those described.

According to these observations, a known protein structure will be a good template for the target protein if the sequence identity between them is = 50%, while it will be generally very difficult to build a reliable model when the sequence identity is lower than 20-30% [26]. It should be mentioned however, that in some cases, even an approximate model based on a sequence identity lower than 20-30% can be useful for many practical applications as long as additional information is available [27].

II. SOURCE OF DATA: PROTEIN SEQUENCE AND STRUCTURE DATABASES

The information required for model building by homology are of two types:

?? 1D information: nucleotide and amino acid sequences;

?? 3D information: protein structures.

These data are stored in databases maintained by groups responsible for collecting, checking, formatting and updating them, for adding annotation, and for making them available to the scientific community [23]. The data can be submitted to the databanks directly by the sequencing groups, extracted from the literature and from patent applications or derived from other databases (for example most of the available protein sequences are obtained through translation of DNA sequences rather than from direct protein sequencing). Most of the databases are cross-referenced to several other databases and appropriate tools are usually provided to retrieve and, in some cases, analyse the data. One of the most important features of protein sequence and structure databases is the possibility to access them directly via the Internet (see URL: `http://www.embl-heidelberg.de/srs/srsc`). Given the growing number of data bases and their continuous improvement, the only reliable font of information about them is the Internet. Here we will just list the most important as a reference point for the reader:

1. Nucleotide (DNA and RNA) sequences are collected by Genbank (NIH, USA) [28], by the EMBL Data Library or Nucleotide Sequence Data Bank (EMBL, Heidelberg, Germany) [29], DDBJ, by the DNA Database of Japan [30].
2. Amino acid sequences are collected by the group at the National Biomedical Research Foundation (Washington DC, USA), who also developed an information retrieval system called PIR (Protein Identification Resource) [31], by the Martinsried Institute for Protein Sequences, Max Planck Institute for Biochemistry (Munchen, Germany), by the Protein Information Database JIPID (Noda, Japan) and by Swissprot at the EMBL (Heidelberg, Germany) [32]
3. The databases of 3D structures are the Protein Data Bank or PDB [20], at the Brookhaven National Laboratory (New York, USA) which contains structures of biological macromolecules (proteins, nucleic acids and carbohydrates) and the Crystallographic Data Centre (Cambridge, UK) [33, 34] devoted to the structures of small molecules, that can be components or ligands of biological macromolecules.

Behind these principal databases there are several derived ones. The ones listed below are just examples:

PROSITE contains protein sequence patterns (common, for example, to a protein family) or sites (diagnostic of a protein function) [35].

BLOCKS contains aligned 'ungapped' (see below) segments of protein sequences corresponding to their most highly conserved regions [36].
DSSP (database of secondary structure assignments) contains information about the secondary structure assignments for each entry in PDB [37].
HSSP (homology-derived structures of proteins) merges sequence and structure information by providing alignments of the sequence of each protein structure in PDB with all its sequence homologues [38].
FSSP (families of structurally similar proteins) contains structural alignments of proteins in PDB [39].

III. METHODOLOGY: MODEL BUILDING

The essential steps of model building by homology are:

1) Identification of the protein(s) of known structure with the highest sequence identity or similarity with the target sequence; optimal alignment between the target and template sequences and modelling of the main-chain of the core
2) Loop prediction
3) Side-chain modelling

After optimal alignment of two sequences, one can measure their sequence identity (by simply counting the number of identical residues found in corresponding positions) or their sequence similarity (by adding the 'similarity' score between each pair of aligned residues). The problem of finding the alignment of two strings of characters that maximises sequence identity or similarity can be formulated in precise mathematical terms and algorithms able to solve this problem are known since a long time [40, 41]. However, this optimal sequence alignment does not necessarily correspond to the optimal superposition between the two protein structures. This is mainly due to the presence of amino acid insertions and deletions between two homologous proteins and to the relatively arbitrary choice of the similarity score.
The probability that insertions and deletions occur among related proteins is not high and, above all, it is not the same in all positions. For example, insertions and deletions are much more frequent at the protein surface, where they only determine local variations of the structure, than in the core of the protein or within secondary structure elements, since in these regions they most likely affect the protein structure and/or function. As in a protein structure there is a limited number of positions in which it is possible to insert or delete residues without altering the protein function, it is more likely to have the insertion or deletions of a contiguous segment in one of these 'neutral' positions rather than the insertion or deletion of the same number of residues in different positions.
The probability for a mutation to occur depends upon the similarity between the exchanged residues. This similarity can be evaluated on the basis of specific criteria: for example, conservative mutations, that is mutations between residues with similar features (chemical-physical properties, dimension, etc.), can be more easily accepted in a protein structure and mutations between residues coded by nucleotide triplets differing for a single base are more likely to occur.
Methods for protein sequence alignment [40, 42] take into account the above factors by assigning a penalty for insertions and deletions which is higher at the beginning of the insertion and lower for subsequent residues, and often use scoring matrices derived from a statistical analysis of patterns of mutations in protein structures which assign a specific cost

for each residue mutation [43, 41, 44, 45]. These algorithms are able, given a penalty for insertions and deletions and a similarity matrix to give the global optimal alignment between two protein sequences and to measure the identity or similarity between them. They are also extended to provide multiple sequence alignment between members of protein families and can also be tailored to search sequence databases for proteins similar to a given sequence [42, 46-48]. In this case, the algorithms have to compare a very high number of sequences and some of them [46] use approximate alignment algorithms to pre-screen the databases.

In some cases the output is a list of sequence alignments between the sequence of the protein used for the search and similar sequences found in the database [46, 48]; in others it is the alignments between segments of the input sequence with those of database sequences which are not interrupted by insertions and deletions ('ungapped' alignments) [47]. The latter can be useful to detect correlations between proteins which have a relatively high local sequence identity but a poor global similarity.

Using database search methods it is possible to select which proteins of known structure are more similar to our target sequence.

When the sequence identity is greater than 50%, it is generally possible to obtain a reliable alignment using any alignment method. When the identity drops to less than 40% automatically generated alignments usually contain errors, which can often be corrected manually on the basis of different criteria. It is usually advisable to build a multiple sequence alignment of as many proteins of the same family as possible, because this can help in assessing the correctness of the alignment. For example, sometimes secondary structure information on at least one of the proteins of the alignment is available, either from X-ray or NMR structure determination, or can be obtained using secondary structure prediction methods. In this case it is possible to verify whether insertions and deletions fall outside of secondary structure elements and, if not, to modify the alignment appropriately. Many protein sequences contain specific patterns of residues which are characteristic of the family they belong to; the residues belonging to these patterns, as well as those involved in protein function (e.g., catalytic residues of enzymes) should be correctly aligned. The multiple sequence alignment of the target with similar sequences will show conserved and variable regions within the family and this can help in aligning distant homologues. Literature and experimental data should also be used to check and refine the alignment.

It is important to highlight that the correct alignment of the 'target' and 'template' sequences is the fundamental step in any homology modelling procedure: errors at this level are the main cause of errors in the final model (URL: `http://PredictionCenter.llnl.gov/`).

The protein with the highest sequence identity with the target is used as a template for modelling the mainchain of the secondary structure elements of the target. If different regions of the target sequence are most similar to different proteins these can be selected as a template for the corresponding regions [24, 49].

Loops are regions connecting secondary structure elements of a protein and are usually located at its surface. Information about loop structures is often important in that, in many cases, loops have an important functional role: thanks to their surface location, loops are often involved in interactions with other proteins or in the catalytic mechanism of enzymes and, in some cases, they constitute the nucleation site for protein folding [50].

The prediction of loop structures is a particularly difficult task since they are much less regular and much more variable than ?-helices and ?-sheets; moreover, insertions and deletions are most likely to occur in loop regions, therefore their structure is often quite different even among closely related proteins. A satisfactory general method for loop prediction has not yet been developed but, in a few cases, the structural analysis of known proteins has allowed to identify heuristic rules and to develop methods for loop prediction. Other methods have been reported to be able to predict loops conformations in the context of a correct structure [51], but none of them has been successful in either of the CASP experiments, possibly because of errors in the rest of the structure [26, 24, 49] (see URL: `http://PredictionCenter.llnl.gov/`).

In some cases, loop conformations can be inherited from the template structure if their length and sequence patterns are conserved. Alternatively, rules based on known sequence-structure relationships, database searching techniques or *ab initio* calculations are used.

The conformation of short - up to 4 residue long - turns, which allow the peptide chain to change direction of 180^{o} [52], is determined by the presence of special residues like Gly and Pro in specific positions of the loop and can therefore be predicted on the basis of the loop sequence [52-60].

In at least one protein family, immunoglobulins, functionally important loops can be predicted quite accurately: the identification of a limited number of 'canonical structures' for five of the six immunoglobulin hypervariable loops (L1, L2, L3, H1, H2) [61-64] and the recognition of the residues responsible for each of these structures allows to predict their conformation with an accuracy within 0,2-1,0Å [65].

The predictive ability of this method has been validated through rigorous 'blind testing' [63]. Recently, recurrent conformations have also been described for the sixth loop (H3) [66-68] and a prediction method for this loop has been developed [68] (Fig. 1). It is now possible to accurately predict the conformation of the 10 residues close to the framework of H3 loops of any length, the overall conformation of H3 loops up to 12 residues in length and, in some cases, the overall conformation of longer H3 loops (Morea, V., Tramontano, A., Rustici, M., Chothia, C. and Lesk, A.M., Conformation of the third hypervariable region in the VH domain of immunoglobulins, submitted)

One of the most commonly used procedures for loop prediction consists in searching the database of known protein structures for regions with a similar conformation to that of the regions adjacent to the target loop and separated by the same number of residues as those of the loop. This procedure is based upon the hypotheses that i) a loop with the same conformation of the target loop is present in the database and that ii) there is a relationship between the conformation of the adjacent regions and that of the loop, that is adjacent regions with a similar conformation are connected by loops with a similar conformation. However, it has been demonstrated [69] that the structural similarity between regions preceding and following loops of the same length is neither a sufficient or a necessary condition for the structural similarity of the loops themselves: similar adjacent regions can be connected by loops with either similar or different structure and structurally similar loops can have similar or completely different adjacent regions. Therefore, while it is possible to identify loops with a similar structure to the target loop using database search

techniques, it is not possible to distinguish *a priori* a correct result from a wrong result, and this limits considerably the utility of the method.

Ab initio methods for loop prediction [70-75] do not use the data base of protein structures. They generate different putative loop conformations and evaluate them on the basis of empirical energy functions, often taking into account the interactions between the modelled loop and the core of the protein structure. However the loop to be predicted can interact with other loops that have to be modelled, so that the complete panoply of interactions cannot be taken into account. The methods used to generate and evaluate the different conformations will be described in the *ab initio* protein structure prediction section.

These methods have two major limitations.

The available force fields (see below) are not sufficiently exact or complete to evaluate correctly the energy of the different conformations. Also, the model of the regions in which the loop is to be inserted is affected by an error which could strongly influence the results of a detailed energy calculation [50].

The evaluation of the energy of different conformations of short protein segments outside their structural context [70-75] does not take into account tertiary interactions, that is interactions with residues outside the loop which have been demonstrated to be important determinants in many cases [76]. Moreover, it has been shown that pentapeptides with the same sequence assume different conformations in different proteins [77]. Consequently, methods which do not take into account the specific environment of the loops will be able to predict only those loops whose conformation is determined by local interactions, while they will give incorrect results in the other cases. On the other hand, loops whose conformation is determined locally can often be predicted from sequence only, in a simpler way than using complex *ab initio* calculations.

In order to produce a complete model, the main chain conformation of the core and those of the loops, however obtained, are merged together. However, as the available methods to predict loop conformations are able to identify the correct conformation just in a limited number of cases [78, 79], it is advisable to critically evaluate whether it is really necessary to model them: while some loops are critical for protein functions, others are far from the regions of main interest of the model (e.g., binding or catalytic sites) and their prediction can therefore be omitted.

The next step in a homology modelling experiment is the assignment of the side chain conformations.

Each amino acid side-chain preferentially assumes a limited number of conformations [80], usually collected in a so-called 'rotamer library'.

These libraries can differ in the grouping of amino acid used to calculate the statistics of the rotamer distribution, for example by taking into account the local environment of a residue or its backbone angles [81, 82]. In some cases, these libraries, combined with some method to exclude rotamers producing unfavourable steric interactions, are used to build all the side-chains of the model. However, as the target and template proteins are assumed to be related, the conformation of the side-chains of the conserved residues of the target can be modelled on that of the corresponding residues of template; the non conserved residues of the target can also be modelled by importing the conformational angles of the template up to where the relative length of the two side-chains permits, using rotamer libraries for the remaining part of the chain.

As for loops modelling, energy-based procedures are also used. Usually these procedures start their refinement from a model having the most common rotamers at every position [82].

While methods for modelling side-chain conformation seem to perform rather well when given experimental co-ordinates for the backbone atoms [49], their accuracy is much lower for protein models and decreases rapidly as the r.m.s.d. between the model and the real structure increases [26] (see URL: `http://PredictionCenter.llnl.gov/`). This might indicate that an improvement in this area could be automatically achieved as a consequence of improvements in backbone modelling [26].

IV. MODEL REFINEMENT

After a complete model has been built, this has to be inspected, both visually and through the use of specific programs, to evaluate and optimise it. Unfavourable steric interactions have to be relieved, by changing the side-chain conformations through few cycles of energy minimisation or geometric refinement. Both techniques can also be used to optimise those main-chain regions that, because of the insertion of loops, result from joining fragments coming from different proteins. However, it should emphasised that neither methods can substantially modify the starting model and consequently adjust large mistakes [23].

Many attempts have been made to obtain a global refinement of a protein model using energy minimisation techniques or molecular dynamics (see: *ab initio* methods) but it has not yet been proven that these methods can improve the quality of the model. Energy minimisation algorithms will only find the local minimum closer to the starting conformation. Energy minimisation of protein crystal structures usually leads to a local minimum with an r.m.s.d. of about 1.0Å from the native structure, which is comparable to the expected error for a model built from a template protein with sequence identity = 50% [25]. In blind tests, energy minimisation and molecular dynamics did not improve the quality of the models and often models built without any further refinement were closer to the real structures than those 'optimised' using various combinations of these methods [26] (see URL: `http://PredictionCenter.llnl.gov/`).

V. EXPECTED ACCURACY OF THE MODEL

The overall quality of models is highly dependent on the quality of the sequence alignment and on the degree of similarity of the target with the parent structure. Both factors are related to the degree of sequence similarity and to the number of insertions and deletions between target and parent [26]: a model built for a target with a medium to high sequence identity (> 40 %) and without insertions or deletions with the template is generally highly accurate [26] and can be almost as accurate as crystal structures when sequence identity is high (~ 85 %) (see URL: `http://PredictionCenter.llnl.gov/`).

An upper threshold for the accuracy of homology models can be established based on the differences between different structural determination of the same protein which have r.m.s.d. values around 0.25-0.40 Å) [25]; a model cannot be expected to be better than this range.

The correctness of the alignment is the main factor influencing the r.m.s.d. between the model and the real structure: even small errors in the alignment give rise to high r.m.s.d. values, while a correct alignment will allow to produce very good models, at least for the core regions, even in predictions based on distantly related parent structures (see URL: `http://PredictionCenter.llnl.gov/`). This emphasises the need for careful analysis

and manual editing of the alignment for pairs of sequences with < 40 % of identity since automated methods do not provide good alignments in this range .

Given a correct alignment, the quality of the prediction of the main-chain in the core region of the target protein can be evaluated on the basis of the relationships between sequence identity and structural similarity previously described [25].

It is worth mentioning that some completely automated methods tested in the CASP experiments [83, 84] proved to be able to build correct models when sequence identity with the parent is very high (85%); however, for more distantly related proteins human intervention, first of all in the correction of the sequence alignment, is still required to obtain reasonably accurate models [26].

The quality of the prediction should be evaluated using our knowledge of protein structures:

?? The determination, through appropriate programs [85], of the solvent accessible surface of each atom or residue allows to distinguish between buried and exposed residues and to assess whether the partition of hydrophobic and hydrophilic residues between the surface and the core are comparable with what is observed in real protein structures.

?? The determination of atomic volumes allows to evaluate the packing and to identify cavities larger than those usually present in protein interiors [86].

?? Unpaired hydrogen bond donors or acceptors should not be present in solvent inaccessible regions of the proteins [26].

?? The stereochemical quality of the model can be evaluated on the basis of standard parameters derived from the statistical analysis of known protein structures [87].

One conclusion derived from the comparison of different predicted structures is that the best models are those which deviate less from the parent structure range (see URL: `http://PredictionCenter.llnl.gov/`). In other words, any attempt to model *ex novo* regions of the protein or more sophisticated approaches which inherit their structures less directly from the parents seem to perform less well: this implies that modelling techniques are still not able to add features to the models. Open problems are the modelling of main-chain segments whose conformation differs from that of the parent structure or that are shifted as rigid bodies with respect to the parent [49], the modelling of loops other than those predictable from sequence, the modelling of side-chains when the backbone conformation of the parent deviates significantly from that of the target.

It is important to note, though, that the most conserved regions in proteins are those that have an important structural and/or functional role and these regions are often those modelled with higher accuracy [23]; therefore, in spite of their possible shortcomings, models built by homology generally contain a wealth of practically useful information and are often instrumental in interpreting experimental data, in planning new experiments and in guiding the design of modified proteins [49].

FOLD RECOGNITION METHODS

Fold recognition techniques try to identify known protein structures which are compatible with a target sequence, even if the template and the target share no detectable sequence similarity. The rational basis for these methods are the following:

1. The relationship between protein sequence identity and structure similarity [25] is not biunivocal: proteins with high sequence identity invariably have similar structures but proteins with similar folds can arise from both similar and completely different sequences (and functions).

2. The majority of known protein structures can be grouped into a limited number of structural classes [88]; it is therefore likely that the number of possible protein folds is limited [89].

As a consequence of these observations, there is a reasonably high probability that the protein structure database contains structures similar to that of a target protein, even if sequence search methods are unable to detect the similarity. This probability will grow as new proteins structures with novel folds will be determined. However, if the target sequence shows no significant sequence identity with proteins of known structure, new criteria to identify the related known protein structure have to be developed.

Because of the reasonable success of fold recognition techniques in CASP1 [90], they have been the field of protein structure prediction which expanded more rapidly: in CASP1, 8 groups participated in this section while in CASP2 the groups were more than 34 [91].

Also in fold recognition, the use of evolutionary information can improve the results [92] (see URL: `http://PredictionCenter.llnl.gov/`), for example a prediction on a given target protein is more likely to be correct if the same prediction is obtained on a distantly related protein.

I. Methods

Several different strategies have been elaborated and used to recognise sequence to structure compatibility. These strategies can be ascribed to one of the following categories: profile based methods, threading and mapping methods.

Profile based methods [93, 94] rely on the observation that each amino acid residue shows preferences for specific structural environment and that consequently some residue types are more likely to be found in a given position in a protein structure than others.

From a statistical analysis of the data base of known protein structure, it is possible to classify each amino acid in classes, for example â given amino acid type can be more often found in buried regions of ? -helices, or in exposed loops.

Given a target protein sequence, each amino acid can be substituted by a symbol representing the class it belongs to. Conversely, each position in a protein structure can be represented by a symbol describing its environment (exposed or buried, ? -helix or?? -strand). These two mono-dimensional strings can then be aligned by applying the same algorithms used for sequence alignment [93, 95] and the quality of the alignment between the structure and the target sequence can be evaluated. The alignment score will be related to the probability that each residue of the target sequence will be found in the environment of the corresponding structural position and will represent the overall compatibility between the target sequence and the 3D structure. Variations of this method combine information about environmental preferences with sequence substitution matrixes.

In threading methods [93, 96, 97, 95, 98-100], the target sequence is inscribed in all possible frames into a subset of the known protein structures, selected to be as representative as possible of the different types of existing folds. The different alignments between the target sequence and each of these structures are evaluated by using some energetic function. The assumption underlying these methods is that the native protein structure corresponds to the lowest energy conformation among those accessible to the protein chain at equilibrium; consequently, alignments with low values of the energy function should be indicative of the compatibility between the target sequence and the 3D structure.

The critical components of these methods are considered to be the energy functions describing protein-solvent systems, the techniques used to perform sequence-to-structure

alignments and the criteria chosen to identify known structures which are similar to the native fold of the target sequence.

A wide variety of threading methods have been developed, using different approaches to face these issues.

Two main strategies have been used up to now to develop energy functions able to describe molecular systems [101]: 'inductive' approaches start from *a priori* chemical-physical principles and use quantum-mechanical calculations to generate semi-empirical energy functions like those used by *ab initio* protein structure prediction methods (see below); 'deductive' approaches, start from the experimental data, that is from known protein structures, and use statistical analysis to generate knowledge-based energy functions. The latter, because of their relative simplicity, are the most commonly used in threading methods.

In knowledge based energy functions, the frequencies of observed events (for example of contacts between two amino acid types) are extracted from the database of known protein structures and transformed in energy terms by applying the inverse Boltzmann's equation [102].

The formulation of the potential energy as a function of inter-residue contacts is based on the assumption that in protein structures pair-wise inter-residue contacts between non-bonded amino acids are determined by the interaction energy between the two residues; in other words, two residues which are often found close to each other are likely to establish attractive interactions. Of course, this assumption is not, or not completely, true. In fact, in proteins, any two residues also establish interactions with several other residues which might be determinant for their relative position. For this reason, more complex potentials which try to take into account the contacts among triplets and 4-tuplets rather than pairs of residues have been developed and reported to improve recognition of native folds (see URL: `http://PredictionCenter.llnl.gov/`).

Several other different formulations of the potential energy function have been proposed both more complex and simpler than the first pair-wise residue potentials [102]. More complex potentials include additional terms behind those accounting for residue-residue interactions, for example solvent accessibility and backbone conformation terms [102]. However, it has been shown that there is no substantial improvement in using all these terms together: even a very simple potential considering only contacts between buried hydrophobic residues (Leu, Ile, Cys, Met, Phe, Trp, Val) proved to be reasonably successful [102].

Current potentials also differ in the representations used for amino acid residues [103]: each residue is often represented by a single atom, for example C? or C?, or by average side-chain centroids; in some cases backbone and side-chain groups are distinguished.

Up to now it has been quite difficult to compare the fold-recognition ability of knowledge-based potentials [90, 91]. A generally used test is the so-called 'self-recognition' test, in which the native fold of a protein sequence has to be recognised by threading the sequence into a library of known folds, with no gaps in the sequence and in the fold allowed in the threading process ('ungapped-threading'). This is a necessary but not sufficient test, in that it is too easy (the native structure is much more favoured with respect to the other alternatives and even simple patterns of hydrophobic and hydrophilic residues have been shown to be able to identify it) and its success does not guarantee the recognition of similar structures when the native fold is not present in the database [102]. New tests are therefore

being developed to evaluate the ability of the potentials to discriminate between real and purposely built decoy structures [91] (see URL: http://PredictionCenter.llnl.gov/) or by requiring the identification of structural homologues in a database of known structures sharing = 25 % sequence identity with each other and from which the native fold has been excluded.

The results obtained from blind tests suggest that current potentials are quite similar in their ability to recognise the native conformation of a target protein [102]; thus, an excessive complication of the potentials does not appear to be justified especially since, as the structure database grows, the speed of the algorithm becomes a relevant issue.

Although recent work demonstrates the theoretical basis of the Boltzmann's formulation [104] , these potentials have been criticised because they would not represent physically realistic force fields: for example, the potentials for equal charges is similar to that derived for opposite charges, probably reflecting the tendency of charged residues to lie on the surface rather than a specific interaction between them [105]. However, as the aim of knowledge-based potentials is to predict protein structures rather than to represent the 'true' physical forces, essentially any potential which works can be considered a useful tool for protein fold recognition [91].

Once an energy function has been defined, the sequence is threaded into a library of structures; the 'threading' consists in 'inscribing' the target sequence into each structure so that each residue of the sequence replaces one residue in the structure. The alignment which provides a low value of the energy function that should correspond to folds compatible with the target sequence.

Different solutions have been proposed for the choice of the library of folds, the way to treat insertion/deletions in the sequence-to-structure alignment, the way to substitute residues from the target sequence into the structures and the algorithm used to align sequences with structures.

Libraries of folds are constructed by selecting non-redundant entries representative of all known folds from the protein structure database. In some cases, just the secondary structure elements in the protein core regions rather than the overall structures are used, based on the rationale that this is the only part of the structure which is conserved among distantly related proteins [92] and references therein. However, as it has been shown that only part of the 'key-residues' responsible for correct fold recognition are found within secondary structure elements [92], this choice could prevent fold recognition.

In some cases, idealised folds rather real ones have been used [92]. They have been obtained, for example, by modifying the topologies of native protein core regions; therefore, even if fold recognition techniques are used, the prediction can be considered as *ab initio*. However, the results obtained for the modified topologies are worse than those obtained for experimental ones, suggesting that the modified topologies lack some crucial features which are necessary to recognise native folds (see URL: http://PredictionCenter.llnl.gov/)

Even among structurally related proteins, insertions and deletions ('gaps') are likely to occur; the way in which gaps are treated varies notably among the different threading approaches.

The insertion of gaps increases the computational complexity, so many threading approaches do not allow gaps at all in the sequence-to-structure alignment ('no-gap' or 'ungapped' threading): these approaches mount the sequence on a portion of structure of

equal length used [92]. 'No-gap' threading is not generally used for prediction purposes in that if insertions and deletions are ignored it will be difficult to find a good alignment of the target sequence with a similar structure [99] ; still, as the native fold is usually recognised, in spite of bad alignments [106], this method is generally used to test the potentials [92].
For prediction purposes, gaps of variable length are often allowed both in the target sequence and in the structures used [92] and references therein; in these cases, variations of loop length and conformation among structurally related proteins would not prevent recognising a similar fold for the target structure. As is the case for sequence alignments, the choice of the penalty associated to insertions and deletions is quite relevant [100].
To reduce computational complexity some approaches substitute the amino acids of the target sequence to the amino acids in the structures one at the time leaving the rest of the structure unmodified; in this way, each residue of the target sequence is surrounded by the residues of the structure onto which it is mounted, rather than by the corresponding residues of the target sequence. This approach, called 'frozen approximation', is quite rough; still, in 'blind tests' it performed as well as more sophisticated methods [91]. The reason probably lies in the fact that the 'frozen approximation' may be appropriate for the recognition of similar folds, provided that conservative substitutions (e.g.: replacement of residues with similar chemical-physical properties) have occurred between the native fold of the target sequence and a known structure with a similar fold; if this is the case, even if the environment of the similar structure is not the same as that of the native one, it will be sufficiently close to allow recognition [91].
For each sequence-to-structure alignment generated by the threading algorithms the value of the energy function is calculated and used to evaluate the likelihood that the sequence can assume a fold similar to that of each structure in the data base [106].
A useful measure of the goodness of the sequence/structure alignment is the z-score, usually calculated by most of the available programs, defined as:

$$z = (E - Em)/?$$

where:

E = energy of the given alignment
Em = average energy over all alignments
? ? = standard deviation

Large negative values of the z-score for the alignment of a sequence with a particular structure indicate that the sequence is likely to be compatible with the structure.
It is useful to calculate the total interaction energy of each residue along the amino acid sequence: native structures generally have energy values below zero in most sequence positions with only few weak positive peaks, and a sequence correctly aligned to a similar fold, does not show many positive values [101].
There are factors that could affect the value of the potential energy function and therefore the results of fold recognition experiments. For example, structures with similar sequence length and/or amino acid composition to the target sequence could be erroneously scored as similar; moreover, the higher is the number of possible alignments (because of the length of the alignment or of the higher number of gaps allowed) the higher the probability that a good alignment can be found by chance. Scoring schemes have been proposed in order to correct for these possible artefacts, and this has indeed been shown to reduce the number of false positives [100].

Fold recognition can also be achieved by comparing sequence-based predictions obtained through 1D and 2D methods, for example predictions of secondary structure, solvent accessibility and long-range contacts (see below), with analogous information extracted from known protein structures (see for example [107-109]).

The development of these methods, called mapping, has been catalysed by the improved accuracy of secondary structure prediction methods which are now accurate enough to serve as the basis for tertiary structure predictions [110]. The accuracy of mapping methods is therefore expected to increase with improvements in *ab initio* prediction methods for secondary structure and solvent accessibility.

Some methods [109] compare the secondary structure assignments (?-helices and ?-strands) predicted for the target sequence from multiple sequence alignments with the secondary structures extracted from a library of protein domains {allowing for insertions and deletions of whole secondary structure elements} to find all possible domains whose secondary structure matches that of the target sequence. A series of 'filters' based on simple rules about protein structures are then applied to these matches (or 'maps') to restrict the number of plausible folds. Among the filters used there are, for example, the observed and expected values for the radius of gyration, the distance between co-ordinates that have to be bridged by loops of a certain length, the ?-sheet topologies (e.g., folds with isolated ?-strands are removed) and distance restraints from experimental data (e.g., NMR measurements, presence of disulphide bridges or of clusters of functional residues). The patterns of predicted and experimental solvent accessibility are used to align the sequence of the target and that of the remaining folds and the final alignments are evaluated on the basis of accessibility and secondary structure matching. This procedure is able to reduce the number of possible folds for a target protein to a few plausible alternatives, and ideally to just one match [109]. The accuracy of this method is reported to be comparable or better to that of the more computationally intensive threading methods to recognise native-like folds and to correctly align amino acid residues and secondary structure elements [109]. As mapping methods heavily rely on the accuracy of secondary structure predictions and these, in turn, have been shown to be much more reliable when based on multiple sequence alignments rather than on a single sequence, it is believed that the essential pre-requisite for successful fold recognition through mapping methods is to start from a high quality multiple sequence alignment containing sufficient number of adequately diverse sequences [107].

II. Model building

When fold recognition procedures identify a significant match between the target sequence and a known structure, this structure can be used as a template to build a model of the target protein. Model building follows the same steps described for homology modelling: the main-chain of the secondary structure elements can be modelled on that of the target protein while for side-chains and loops prediction alternative strategies have to be used. Once again, the quality of the alignment will be the main parameter in determining the quality of the final model.

III. Accuracy

The assessment of the performance of fold recognition methods is not a straightforward task in itself, in that methods to compare structures of unrelated proteins [90] and criteria to decide if such structures are similar or not have to be developed [111]: in some cases predicted and target structures are almost identical while in other cases the similarity is borderline and often several possible alignments can be obtained [90].

The criteria used in the first protein structure prediction assessment experiment were not very stringent (and were in fact modified in CASP2):

1. The fold was considered as correct if a significant fraction of the secondary structure elements of the selected fold could be aligned with the target structure with an r.m.s.d. = 3.0 Å. Moreover, both the best (i.e., the lowest energy) hit and the first 10 best hits were taken into account; the reason to consider the first 10 hits is that generally they all have very similar scores and even correctly recognised folds do not have a score significantly higher than incorrect folds [90].
2. The secondary structure segments were considered as correctly aligned if at least one residue of a predicted element overlapped with the correspondent secondary structure element in the target; other indicators of the goodness of the alignment were the number of residues by which secondary segments were shifted along the sequence and the average shift over the whole sequence and the number of correctly aligned secondary structure segments versus the number of theoretically alignable elements [90].

In CASP1 [90] the success of fold recognition methods in rigorous 'blind tests' was partial; however, fold recognition is the method for protein structure prediction that has shown the biggest improvement in the second experiment [92].

In both experiments each of the methods proved capable to identify some of the folds, in the absence of detectable sequence homology between the target and a protein of similar structure even in cases where the similarity between the target structure and the known folds was rather low [102, 91]. However, although all the targets were recognised by at least one method [111], no method was able to recognise all the targets, even if the number of targets identified by each group increased in the second competition [92]. Moreover, clear methods to assess the correctness of a result *a priori* are still lacking [90].

Folds which are more represented in the protein structure database are identified more easily (usually the prediction is correct), probably because the potentials derived from the structure database are biased in favour of these structures, and also because the libraries used for fold recognition often contain just a copy of less common folds and several copies of the more frequent folds, therefore there is a higher probability to identify the latter by chance [90].

The quality of the alignments of the target sequence to correctly recognised folds, provided by threading methods, is correct in just a small number of cases and it is often distant from the optimal alignment: it is generally more difficult to obtain a correct alignment than to recognise a correct fold [102, 91].

1. In the first competition significant local shifts in secondary segments were observed [90, 102, 91]. This is quite a serious limit in that the lack of an accurate alignment prevents the construction of a useful 3D model for the target protein. For example, the predictions submitted for the core regions were not good enough to allow a reliable modelling of the loops [90]. In the second meeting, accurate alignments have been provided only for targets that, despite the lack of significant sequence identity with known proteins, could be easily recognised as homologous to some known folds based on the similarity of their function and on the presence of conserved key-residues. Nevertheless, alignments provided by fold recognition methods were considerably better than those obtained with sequence alignment methods, which means that these methods could be used to align protein sequences with very low homology [91]

The two CASP experiments demonstrated that the prediction results could be positively affected by human intervention: manual adjustments of sequence alignments, visual inspection of the selected fold, comparison of the secondary structure of the selected fold with the secondary structure predicted for the target and consideration of common functions between the target and the fold [107, 90, 91]. As an example, two of the participants to CASP2 were able to identify the correct fold for some targets based on just the predicted secondary structure of the targets and their deep knowledge of protein structures and their relationships with the function [91]. However, manual intervention is not always successful; in some cases, correct automated predictions have been discarded in favour of worse alternatives [90].

Finally there is a distinction between 'strong' and 'weak' fold recognition [102]: strong fold recognition attempts to find the known fold which is structurally most similar to that of the target protein while weak fold recognition attempts to identify a small set of folds which are compatible with the target sequence and that could be subsequently analysed, for example considering similarity in function or experimental constraints. Weak fold recognition is probably a more realistic goal to achieve and is potentially able to provide very good results when combined with other information [102].

AB INITIO METHODS

Unlike homology modelling and fold recognition methods, *ab initio* methods for protein structure prediction do not use proteins of known structure as templates. However, also these methods use information contained in known protein structures, although less explicitly, to understand general principles governing protein architecture, to derive forcefield parameters, or as input for neural network systems. Some methods also join knowledge about protein structures to *a priori* chemical-physical principles. *Ab initio* methods could be used both to predict whole protein structures or parts of them, although the former is still beyond the capabilities of the existing algorithms.

I. CLASSIFICATION OF *AB INITIO* METHODS

Two strategies are usually applied to predict *ab initio* the tertiary structure of proteins: the first one (primary -> secondary -> tertiary) consists of two steps: the prediction of the secondary structure from the amino acid sequence (primary -> secondary) and the assembly of the secondary structure elements in a 3D structure (secondary -> tertiary); the second consists in the prediction of the tertiary structure directly from the sequence (primary -> tertiary) [112].

A great variety of methods are usually ascribed to the *ab initio* category, and they can been classified, according to the type of information that they can give, in [91]:

?? 0D methods: predict which fold class a protein is most likely to belong to (all alpha-helix, all beta-sheet, alpha/beta or alpha+beta [113-115]

?? 1D methods: predict secondary structure elements (alpha-helix, beta-strand, loop) and residue accessibility

?? 2D methods: provide prediction of long-range contacts (within whole elements of secondary structure or single residues)

?? 3D methods: provide predictions of tertiary structures (overall fold or shape of the protein or 3D co-ordinates)

Alternatively, *ab initio* methods can be classified, on the basis of the principles they rely on [23] as:

?? Methods based on conformational energy calculations, e.g.: search for the most stable protein conformation, protein folding simulations.

?? Methods based on the variability in families of aligned sequences, e.g.: secondary structure prediction, prediction of long-range contacts, prediction of functional residues. These methods are based on the observation that significantly more information is contained in the evolutionary history of a protein. The starting point for all these methods is therefore a good multiple alignment of the target sequence with sequences of homologous proteins.

II. SECONDARY STRUCTURE AND SOLVENT ACCESSIBILITY PREDICTIONS

II.a. Methods

As most of the residues in protein structures are part of regular secondary structure elements (alpha-helices, beta-strands, reverse turns), a lot of effort has been devoted to obtain accurate predictions of these segments [60]. This would be an important step toward the complete 3D structure prediction, provided that a reliable method to assemble correctly these element in space (for example based on the prediction of long-range interactions) would become available.

Several methods of secondary structure prediction are based on statistical information derived from the analysis of known protein structures and sequences.

Amino acid residues show conformational preferences for secondary structure elements and for specific positions within these elements (alpha-helix, beta-strand, reverse turn) [58]; even though these preferences are not very strong, the clustering in the sequence of several residues preferring one type of secondary structure suggests the presence of that secondary structure [60].

The periodicity of hydrophobic and hydrophilic residues can be typical of specific secondary structures: alternate hydrophobic and hydrophilic side-chains are likely to be part of a strand in a beta-sheet with an hydrophilic face exposed to solvent; alpha-helices should have an hydrophobic residues every three or four residues to allow an hydrophobic face to pack against the rest of the protein [116].

Among the methods for the prediction of regular elements of secondary structure (alpha-helices and beta-strands) described in the literature [117, 118], those exploiting the evolutionary information contained in a multiple sequence provide the best results [112, 118]. Secondary structure elements are usually conserved in homologous protein structures and therefore a consensus prediction obtained for all sequences of the alignment is likely to be more accurate and reliable than a prediction performed on a single sequence because [118]. The PHD program which uses multiple sequence alignments as input to a neural network system provides particularly good results: the estimate of the accuracy is slightly higher than 70% [118]. Other algorithms based on multiple sequence alignments give results almost as good [119-121].

The prediction of solvent accessibility can be useful to predict the spatial orientation of secondary structure segments. Solvent accessibility has been described in terms of two (buried/exposed) [122-124], three (buried/intermediate/exposed) [125, 126] and ten states [127].

A system of neural network, analogous to that used to predict secondary structure, has been developed for the prediction of solvent accessibility [127]. This system gave better results than other methods thanks to the additional information contained in multiple

alignments and to the usage of a ten-state rather than a three-state model for relative accessibility [127].

An intrinsical limitation of these methods is that solvent accessibility is much less conserved than secondary structure in homologous proteins, therefore the information that can be extracted from multiple sequence alignments is lower.

II.b. Accuracy

The accuracy of secondary structure prediction methods depends on several parameters, for example, the number and type of sequences in the multiple alignment: when the sequences are few or the degree of diversity between them is small, the quality of the prediction is lower [112]. The multiple sequence alignment itself can be modified on the basis of 'expert' knowledge; human intervention and the use of statistical methods [128] has been reported to improve the prediction accuracy in a number of cases [129, 112, 107, 109].

Blind tests of the PHD program have confirmed the claimed accuracy ($Q_3 = 72 \pm 9\%$)[2] [118] and have shown that the predicted reliability correlates with the observed accuracy. Secondary structures not present in all the families of the alignment and the ends of secondary structure elements are often difficult to predict [112].

The level of accuracy reached by these methods is good enough for other methods (e.g., fold recognition methods) to use the predicted secondary structures elements, possibly together with other restraints [117], as a useful starting point to build a 3D model or as criteria to assess the reliability of the results.

For methods relying on multiple sequence alignment one limitation is that the prediction of secondary structure elements is less accurate for those elements which are not common to all the family. On the other hand, elements which are in the core of the protein are those less likely to diverge even in distantly related proteins [25]; this makes the identification of the correct tertiary fold possible even if secondary structure elements outside of the core are not predicted correctly [129].

Secondary structure predictions, similarly to tertiary structure predictions, seem to be less efficient for unusual folds [24]; possibly, in both cases, because of the bias present in the parameters derived from the current database.

III. PREDICTION OF LONG-RANGE INTERACTIONS

These methods aim at predicting the relative spatial position of predicted secondary structure elements or residues either by predicting long range interactions or by using combinatorial approaches and/or semi-empirical rules

III.a. Methods

The term 'long-range interactions' is used to described interactions among residues which are spatially close to each other in a 3D protein structure but far from each other in the protein sequence. The methods developed to predict long-range interactions exploit information contained in single sequences or, more often, in multiple sequence alignments, to give matrices of predicted inter-residue contacts in a protein (2D prediction), called 'contact maps'.

One of these methods [131] predicts long-range interactions between ?-strand residues in ?-sheets and is based on the statistically derived frequencies of pair-wise inter-residue contacts. Residue distribution on adjacent ?-strands has been shown to deviate significantly from randomness so that pair-wise preferences could be extracted from known protein

---

[2] Q_3 indicates the percentage of residues correctly predicted to be in one of three states: helix, strand, other [130].

structures [132]. In some cases, these preferences can be rationalised on the basis of complementary chemical-physical properties between directly interacting residues (e.g., Ser/Thr and Val/Ile are favoured, Thr/Val and Lys-Arg/Leu are not). In other cases, it is more difficult to rationalise the observed preferences, possibly because the interaction between residues is mediated by solvent molecules at the protein surface or by the packing environment in the protein interior. Specific inter-residue preferences depend upon the ?-sheet topology (parallel or anti-parallel), by the presence or absence of hydrogen-bonds between the backbone atoms of the two residues in contact and by the relative position of the two residues with respect to the N-terminal and C-terminal end of the protein [131]. Based on these preferences, it is possible to recognise the correct pairing of ?-strands in a ?-sheet in known structures with an accuracy of 75% or better. While in principle a similar analysis could be performed for helix-helix and helix-strand interactions, in these cases the lack of strong hydrogen-bonding distance constraints could make the recognition of specific residue-residue contacts more difficult.

Other methods for long range interaction prediction [133-135] are based on the observation that residues in physical contact in the 3D structure in some cases show a correlated mutational behaviour, which can be recognised in a multiple sequence alignment [136, 135]: sequence mutations that could interfere with the maintenance of structure or function within protein families, might be compensated by complementary mutations in nearby positions to allow for the protein (and cell) survival [136]. As an example, if a bulky side chain in the protein interior is substituted by a small one, other residues could mutate appropriately to fill the newly formed cavity. Consequently, it is possible that, if in a multiple sequence alignment two positions mutate in a correlated manner, the residues occupying those positions are in physical contact in the 3D structure. Pairs of residues that are correlated, in the sense described above, do have a weak tendency to have smaller distances in the 3D structure [133] and the method might therefore be useful to predict long-range inter-residue contacts which, in turn, can be of help in modelling the relative spatial orientation of secondary structure elements. One of the major shortcomings of these methods is that the compensatory response of a protein structure to a point mutation is not generally the mutation of another single residue but could involve a cluster of residues [133]; in some cases, the compensatory response to single point mutations is even achieved through small shifts of secondary structure elements [131]. To partially account for this, in some cases clusters of correlated residues have been considered [133].

As residues involved in protein function are generally conserved within protein families, even among distant relatives, the analysis of patterns of conserved and variable residues in multiple sequence alignments can be used to predict functional residues. When the conservation patterns are clear, functional residues can be recognised by visual inspection of multiple alignments, but more subtle patterns of conservation can only be catched through the use of specific tools [137]. The relative spatial arrangement of functional residues can be useful to orient secondary structure elements and to decide about the likelihood of a predicted structure.

The prediction of long-range interactions can also be used to select the correct fold between the candidates generated by fold recognition methods [131] (R, Leplae, T. Hubbard and A. Tramontano, submitted for publication)

III.b. Accuracy

Methods predicting the spatial arrangement of secondary structure elements have been successful in 'blind' tests in recognising folding motifs similar to those of already known structures (e.g., leucine zippers) [129, 107, 135]; on the other hand, unusual folding motifs are still difficult to be predicted [24].

The accuracy of the prediction of ?-strands residue-residue contacts is low when contact maps are generated from a single protein sequence but it can be considerably increased by using multiple sequence alignments and, in some cases, by knowledge-based considerations (e.g., an incorrect prediction of a parallel sheet can be easily recognised if the ?-strands are joined by less than 10 residues: only anti-parallel strands are connected by segments of that length).

The accuracy of the prediction of long-range contacts based on correlated mutations has been reported to be up to five folds better than random [133], and to increase when other information is used. The evolutionary distance between sequences showing simultaneous variations, the specific type of co-variation observed (e.g., volume, hydrogen bonding, charge) and the tertiary structural context (interior or surface) of the co-varying residues can all be effectively taken into account [135]. Restricting the analysis to the residues which are expected to be in the protein interior can also improve the results [135].

Some authors [91] believe that an alignment of a very high number of sequences is required to even attempt a prediction of long-range contacts, and that even if the prediction of such contacts is possible, it is of limited usefulness because of the high number of false positives. Further improvements could be obtained if tools to distinguish between correlated mutations and mutations which do not need to be compensated ('neutral' mutations) were developed.

IV. ENERGY BASED METHODS

These methods try to predict tertiary structure from the amino acid sequence alone that is to solve the classical 'folding problem': given a protein sequence and a model of the interactions between residues, they try to recognise the protein conformation corresponding to its native structure [138]. Although they pursue a very difficult goal, these approaches have the advantage of not depending on the existence of a fold similar to that of the target protein in the protein database and therefore to be potentially able to predict completely new folds. Of course, if the number of protein folds is really limited, as most of the protein folds become known, the need for *ab initio* methods will decrease [112].

These methods consists essentially of two steps: the generation of multiple possible conformations for the target protein and the energetic evaluation of these conformations

Some energy based methods are based on the assumption that the conformation with the lowest energy corresponds to the native state (that is: the folding process is under thermodynamic control). These methods try to generate as many as possible conformations of a protein structure or of a region and to evaluate them on the basis of the energy calculated for each conformation.

Other methods are based on the assumption that a polypeptide chain reaches the native conformation through an energetically accessible pathway, without having to search the complete conformational space. They try to simulate the folding process by dividing it into several steps; at each step they generate and evaluate different conformations, selecting the conformation(s) with the lowest energy as the starting point for the next step [139].

Several methods are available to explore the conformational space of a molecule by varying either its Cartesian or internal co-ordinates (i.e., dihedral angles): systematic approaches, molecular dynamics, distance geometry and genetic algorithms. These methods

Systematic methods evenly explore the conformational space of a molecule by varying its rotable dihedral angles with a pre-established increment, so that all the possible combinations of the selected dihedral angle values are generated. While in principle they can guarantee to sample quite completely the conformational space available to a molecule, their efficiency is limited by the number of dihedrals in the molecule and by the value of the increment for each dihedral. For this reason, systematic searches are usually applied to small protein regions (e.g., loops).

Monte Carlo methods search the conformational space of a molecule through a random or pseudo-random variation of either its Cartesian co-ordinates or, more often, its dihedral angles: in this case, both the increment value and the number of rotable dihedrals varied at each step can be chosen randomly. At any given point of a random search the probability of finding new conformers is proportional to the number of conformers not yet discovered, therefore this probability decreases with the search progress; consequently, also the random methods can adequately cover the conformational space if run for a sufficiently long time.

In molecular dynamics, proteins, possibly together with explicit solvent molecules, are treated on the basis of the principles of classical Newtonian mechanics. As this method is considered to reliably reproduce the motion of a polypeptide chain as a function of time, starting from a random structure and generating a long enough trajectory, the native conformation should be found. However, current computational power is only able to generate molecular dynamics trajectories for time periods considerably shorter (about 10^{-8} s) than *in vitro* folding (typically about 1 s), so the complete conformational space of a protein molecule cannot be explored [60].

Distance geometry is a method to convert a set of distance constraints in a random set of 3D co-ordinates consistent with the constraints: the conformational space of a molecule is described by a matrix of distance constraints including the maximum allowed distance (upper limit) and the minimum allowed distance (lower limit) between any pair of atoms; all the randomly generated conformers lie within these upper and lower limits. This approach samples quickly and efficiently the 3D space but it cannot guarantee that it has been thoroughly searched.

Genetic algorithms are based on mechanisms of natural genes evolution, like mutations and cross-linking: several searches (mutations) are run simultaneously and information is exchanged between them (cross-overs), thus increasing the efficiency of the overall process. These methods have been used both to attempt the prediction of whole proteins and to build loops regions or to find the correct set of side-chain rotamers given the experimental backbone conformation [139].

To speed up the simulations, often these search methods use simplified representations of the polypeptide chain (for example, the amino acid side chains can be represented by spheres and the main chain by only C? atoms) [112]. Moreover, often the search is performed on lattice models rather than in the complete conformational space of the protein molecule.

The only way to calculate the exact energy of a molecule is to use precise quanto-mechanical calculations. As these calculations are computationally intractable for a molecule as complex as a protein, approximate functions are generally used to calculate the

potential energy of the structures generated by the conformational search methods. The form of the energy function can vary, but usually it is the sum of different energetic terms chosen on the basis of the forces that are expected to act on protein structures. As an example, the following energy function takes into account the contribution to the total energy of covalent bonds (stretching, bending and dihedral energy) and non-covalent interactions (van der Waals, hydrogen bond, electrostatic energy):

$E_{total} = E_{stretching} + E_{bending} + E_{torsion} + E_{van\ der\ Waals} + E_{electrostatic} + E_{hydrogen\text{-}bonding}$

The energetic contribution of each of these terms to the total energy is calculated as a function of the deviation of the observed values from a set of previously determined 'ideal' parameters. They represent, for each atom type, the preferred equilibrium positions (for example: length of the N-C? bond, distance between hydrogen bond partners, etc.). These 'ideal' parameters together with the energy function and a set of force constants which penalise the deviations from the 'ideal' parameters constitute the so-called 'forcefield' of a molecule.

Many different forms of this function have been developed. They can contain just a few or only one term or many additional terms (which penalise deviations from planarity of specific groups, strengthen the chirality of specific centres or couple the different energetic terms [23]); in several tests, extremely simple functions result as effective as more complex ones (see URL: `http://PredictionCenter .llnl.gov/`).

Forcefield parameters can be determined in one or more of the following ways: performing quantum mechanical calculations on simple model systems, deriving them from the statistical analysis of known protein structures or measuring them experimentally. Methods which only use quanto-mechanical calculations can be considered as *ab initio* in a strict sense, in that they essentially rely upon *a priori* chemical-physical principles [112]. However, most of the available forcefields do contain empirical parameters extracted from protein structures and are usually tested by evaluating their ability to reproduce known protein structure (see URL: `http://PredictionCenter.llnl.gov/`). A World Wide Web server has been designed to enable an objective evaluation of forcefields and to address important questions concerning forcefield development and application (see URL: `http://iris4.carb.nist.gov/`).

IV.a. ACCURACY

The available *ab initio* methods cannot provide accurate 3D structure predictions yet. With existing methods, just the structure of extremely small proteins for which an extensive conformational searching is feasible can be predicted; however, the predicted structures are still more than 4.0Å from the experimental structures [112]. The main reason why energy based methods have not been very successful is that energy functions and forcefield parameters are neither sufficiently exact or complete to evaluate correctly the energy of the different conformations generated by the conformational search methods. This statement is corroborated by the fact that the energy minimisation of a protein structure experimentally determined gives a local minimum conformation with an r.m.s.d. value of about 1.0Å from the starting conformation; this can be considered a sort of 'resolution' of the forcefield used [23]. Furthermore, because of the errors in energetic parameters, the energy native fold is not significantly lower than that of some of the incorrect folds [140].

It has been suggested that the accuracy of these methods could be improved by exploiting information contained in a multiple alignment; these information could be used, for example, to calculate contact potentials taking into account the variability in a family of aligned sequences [112]. Moreover, as the calculations required by *ab initio* methods are computationally intensive, in this field more than in others an increase in the computer power could allow significant progresses [141-143, 75, 139].

Modelling by homology

I. The background

There is a clear relationship between sequence identity and structure similarity in proteins: the similarity of the backbone conformation in the core regions of two proteins increases with the sequence identity between them (Crippen 1977; Chothia and Lesk 1986; Hilbert, et al. 1993) and this relationship forms the basis for modelling by homology or comparative modelling. This is the most used and also the most effective method to obtain a structure prediction of a protein when its sequence is clearly related to that of a protein of known structure.

It has been estimated that, if applied to all possible targets in the present sequence database, modelling by homology would allow the prediction of at least one order of magnitude more proteins than the protein structures experimentally determined so far (Sali 1995).

I.a. Why build a model

Although structural data on proteins by x-ray and NMR techniques are being produced at an impressive pace and the rate of structure deposition is continuously increasing (Fig. 5.1), the attention devoted to protein

structure prediction is not at all diminishing, and there are very good reasons to pursue this elusive problem.

First, in many cases it can be essential to gain structural information on the protein under study as soon as possible during a project so that effective experiments can be planned (Sollazzo, et al. 1990; Savino, et al. 1993; Orlandini, et al. 1994; Pizzi, et al. 1994; Amati, et al. 1995; Ammendola, et al. 1995; De Francesco, et al. 1996; Luo, et al. 1996; Jackson, et al. 1997; Starling, et al. 1997). Second, the information provided by a model can be instrumental for the experimental determination of the protein's structure (Turkenburg and Dodson 1996). Furthermore, sometimes a model can be effectively used to modify the properties of a given protein or to explain functional differences (Scarborough and Dunn 1994; Failla, et al. 1996).

Last but not least, the challenge of understanding the rules underlying protein folding is intellectually attractive and has been one of the most actively pursued research fields in structural molecular biology, since the early success of Pauling in predicting the structure of an ? helix before any structural determination of a macromolecule (Pauling, et al. 1951).

Although we have witnessed many achievements since, the problem is far from being solved, and in many occasions bursts of enthusiasm about one method or another have been followed by depressing disillusions.

Nevertheless many methods and servers can provide putative answers to the problem of predicting a protein structure particularly when homology

modelling can be used and the aim of this chapter is to highlight the problems underlying the various approaches, to allow the reader to critically evaluate the reliability of the results.

There are three important points to keep in mind in any protein structure prediction experiment. Stating that a model is not an experimental structure is a truism, but it is important to remember that even when there are reasons to trust the results of a modelling experiment, the level of accuracy of any model is, in the majority of the cases, not comparable to that of an x-ray or an NMR structure, even if the pictures of the model are equally colourful and as esthetically pleasant as those of an experimental structure!

Another important point to highlight is that not all parts of the model are equally reliable. As we will illustrate here, each of the steps of the modelling procedure will introduce errors and these errors are not equally distributed over the model. A scenario where a theoretician produces a model that is delivered to the scientist who will use it, is therefore far from being ideal. It is essential that any conclusion derived from a model is carefully checked against the expected reliability of the part(s) of the model involved and that the modeller and the end user of the model work in strict collaboration.

Furthermore the reliability of a model dictates its proper usage. Especially in the case of homology modelling, it is possible to evaluate a priori the reliability of the model, which mainly depends upon the

sequence similarity between the target and template proteins. It is unreasonable to use almost any model for detailed energetic calculations, to predict the binding affinity of a ligand or to design an appropriate ligand. A model can, in most cases, be used to predict which mutations are likely to be accepted by the protein structure, but precise accessibility values of a side chain or packing details can only be derived from a model based on a sufficiently high sequence similarity.

Finally, with the possible exception of models based on very high sequence identity, there is only one way to gain confidence in a model and that is to use it to predict features of the protein. A correct model is one that can be successfully used to modify the properties of the molecule in a predictable way and only well designed and carefully performed experiments can be used to judge the quality of a model.

I.b. Modelling by hand and development of 'automatic' modelling programs.

The observations in the previous paragraph have implications for the usage of servers or programs that perform all the model building steps automatically (Mandal and Linthicum 1993; Sali and Blundell 1993; Srinivasan and Blundell 1993; May and Blundell 1994; Taylor 1994; Peitsch 1996; Peitsch, et al. 1996; Peitsch 1997). Most of these systems have an effective user interface; some have been carefully tested and their output reports the expected accuracy for the various parts of the model. In

the best cases, it is also possible to retrieve the intermediate steps of the model building procedure.

It is essential to use only servers and programs that make the results of their evaluation available to the users, that have been tested in blind trials and to check the results of all the intermediate steps of the automatic model building, with particular attention to the alignment.

II. The steps involved

Model building is a multistep procedure and can be outlined as:

?? identification of the protein(s) to be used as template

?? identification of the regions that are structurally conserved between target and template

?? construction of a sequence alignment in these regions

?? model building of structurally conserved regions (using the coordinates of the backbone of the template)

?? model building of structurally variable regions (where for example insertions and deletions are present)

?? construction of the side chains of the model

?? refinement of the model.

II.a. Selection of the parent structures

Homology modelling relies on the observation that similarity in sequence implies structural similarity so that the co-ordinates of the backbone of

the residues of the template proteins can be used to build a model of the corresponding ones in the target protein. The similarity of the backbone conformation in the core regions of two proteins increases with the sequence identity between them. The core region here is defined by superimposing the backbone atoms of the secondary structure elements, and extending these elements to include additional residues at their ends, as long as the r.m.s.d. is lower than some threshold, usually 3.0 Å. The percentage of residues in this region depends upon the structural similarity between two proteins and, consequently, upon their sequence similarity.

It has been shown that (Chothia and Lesk 1986):

For proteins with sequence identity = 50% the r.m.s.d. of the backbone atoms of the core region is = 1.0Å and this region comprises about 90% of their structure.

For proteins with sequence identity = 20% the core region may comprise as little as 50% of their structure with an r.m.s.d. of the backbone atoms in this region > 1.8Å; significant structural differences can occur outside the core.

Proteins with sequence identity between 20% and 50% have an intermediate degree of similarity between those described.

This implies that the best template for model building is the one sharing the highest sequence similarity with the target protein.

When more proteins of known structure with roughly the same sequence similarity to the target are available it is advisable to select the template according to appropriate criteria, for example resolution or completeness of the structures or state of ligation. A typical case is for example that of immunoglobulins, where there are so many structures and the level of sequence conservation of the core is so high that such elements can and should be taken into account (Tramontano 1995).

However, since sequence similarity is not constant along the whole sequence, the availability of more than one template could also allow us, at least in principle, to use fragments of each to build the complete model by selecting, for each region, the protein sharing the highest 'local' sequence similarity (Sutcliffe, et al. 1987; Sali and Blundell 1993). It is not yet clear how much better such approaches perform, since they have been applied to a limited number of cases, but it seems that they can provide a somewhat better model when the sequence identity between target and template is low, while they can be ineffective in cases where sequence similarity between target and template is high (Martin, et al. 1997).

II.b. Alignment and alignment reliability

We have assumed that one can easily measure the sequence similarity between two sequences (the target and the template in this case) and there are a number of automatic programs that can perform the alignments and measure sequence identity (by simply counting the number of identical residues found in corresponding positions) or their sequence similarity

(by summing the 'similarity' scores between each pair of aligned residues). Usually these algorithms are used in data base searches so that a measure of sequence identity or similarity sufficiently accurate to evaluate the appropriateness of a template is directly reported by these programs.

Deciding which alignment should be used for model building however presents a different problem.

As described in Chapter 4, the problem of finding the alignment of two strings of characters that maximises sequence identity or similarity can be formulated in precise mathematical terms and algorithms able to solve this problem have been known for a long time. However the alignment that maximises sequence identity or similarity is not necessarily that corresponding to the best structural superposition of the proteins, while in a model building experiment one needs to align the amino acids that are in equivalent structural positions in the two proteins.

The importance of a correct initial alignment cannot be emphasised too much. Every step in the model building process will be based on the alignment and even minor errors at this stage will produce major errors in the final model: a shift of the alignment of one residue with respect to the correct one can produce very large errors in the final model (Martin, et al. 1997).

Most of the uncertainties and errors in the alignment are due to the presence of amino acid insertions and deletions between two homologous

proteins and to the relatively arbitrary choice of the similarity score and this has been described in detail before. However it is important to remember that there is a conceptual difference between building an optimal sequence alignment and obtaining an alignment that represents a structural similarity.

For example the alignment shown here, although correct in terms of maximising sequence identity, cannot represent a structural alignment:

```
sequence 1 LADGTRCTGRGSDW
sequence 2 LVD-SKCRAKG-DW
             * *     * *
```

The amino acids marked by stars cannot correspond to each other in the structure, since it is impossible that two residues are in the same position in space and yet they are separated by one amino acid in one protein and none in the other! This is a trivial example, but it stresses the concept that a sequence alignment is a one dimensional result that should be translated into three dimensions.

A more correct way to write the sequence alignment shown above could be:

```
sequence 1 LADGT--RCTGR--GSDW
sequence 2 LV---DSKCRAKGD---W
```

to show that there is no one to one correspondence between the residues contiguous to the insertions: it is important to always try and imagine the structural implications of a sequence alignment.

The CASP2 experiment provides a more realistic example of what we just described (Samudrala and Moult 1997).

One of the target proteins for comparative modelling was Endoglucanase I (t0028), sharing 47% identity with a known structure (1celA).

Automatic alignment in one case provided the following alignment for the region spanning residues 49-70 of the target (Samudrala and Moult 1997):

```
TARGET: CTVNGGVNTTLCPDEATCGKNC
        |         |||||  |  |||
PARENT: CYDGNTWSSTLCPDNETCAKNC
```

while the correct alignment turned out to be:

```
TARGET: CTVNGGV----NTTLCPDEATCGKNC
         |                  |     ||
PARENT: CYDGNTWSSTLCP---DNETCAK-NC
```

In other words, these two protein regions cannot be structurally aligned (the main chain varies by more than 4.0Å), although the sequence alignment produces a convincing result by maximising the number of identical or similar residues.

In an homology modelling experiment, the structure of one protein is known and this information should not be ignored. Once a sequence alignment has been obtained, it is extremely important to view the alignment in the context of the structure of the template protein.

Insertions and deletions are much more frequent at the protein surface, where they only determine local variations of the structure, than in the

core of the protein or within secondary structure elements, since in these regions they are most likely to affect the protein's structure and/or function. Consequently, secondary structure elements of the template should not be interrupted by gaps in the alignment. Similarly, completely buried side chains in the template should not be aligned to charged amino acids in the target sequence. This implies that each alignment, however derived, should be carefully checked and corrected manually. This has been shown to improve the final result in many cases and provides the further advantage of highlighting regions of ambiguous alignment that can be built with a lower level of confidence.

In summary, the sequence alignment should be built taking into account the sequences of the two proteins and the structure of the template. However very often a multiple alignment of the protein family is available, and this case is becoming ever more frequent, given the continuously increasing number of experimentally determined sequences.

Such a multiple alignment contains information about conservation and variability at each position and can be of great help in defining the positions of insertions and deletions (Altschul 1989; Henneke 1989; Subbiah and Harrison 1989; Thompson, et al. 1994a; Thompson, et al. 1994b).

In a field as difficult and challenging as protein structure prediction, one cannot afford to discard any information however it is obtained. Another

important aspect is that any experimental information on the protein(s) has to be taken into account. Besides the obvious importance of locating and aligning correctly the active site residues of enzymes, there is often plenty of data on site specific mutants, location of epitopes and so on. Because of the relevance of the alignment for all subsequent steps, this is the stage where all the available information should be collected and taken into account. The alignment should be consistent with all that is known about the protein family under study.

II.c. Structurally diverse regions: can we recognise them

The analysis of a (multiple) sequence alignment can also be useful to highlight which regions of the target protein are more likely to be 'structurally divergent', that is which regions do not structurally correspond to any of the template's regions. The local sequence similarity, the conservation pattern, the location of the region in the template structure, the absence of insertions and deletions, are all elements that can allow us to predict that a region is part of the structurally conserved core and can be built using the backbone of the template.

If more than one structure of members of the template family are available, they can be used to derive the extent of the conserved core of the family. Structural superposition of the available proteins of known structures will highlight which regions tend to be conserved during evolution and which are subject to variation and refolding (Kanaoka, et al. 1989; Taylor and Orengo 1989b; Zuker and Somorjai 1989; Orengo and

Taylor 1990; Rippmann and Taylor 1991; Orengo, et al. 1992; Russell and Barton 1992; Gracy, et al. 1993; Grindley, et al. 1993; Luo, et al. 1993; Orengo, et al. 1993; Rufino and Blundell 1994; Taylor, et al. 1994; Holm and Sander 1995a; Gotoh 1996; Holm and Sander 1996b; Koch, et al. 1996; Koch and Lengauer 1997; Schmidt, et al. 1997).

Even after the alignment has been built and the model building steps have started, it is still worth re-examining the alignment when problems are encountered. For example, if some regions appear difficult or impossible to model, because the resulting packing of side chains is either too tight or too loose or the number of residues in a loop is inconsistent with the distance between the neighbouring regions, the alignment should be checked again, using the new insights derived by the attempts to build the model.

II.d. The loops

As discussed, structurally variable regions cannot be built using the template, and some other method has to be employed. In general, these regions correspond to loops, that is regions connecting secondary structure elements of a protein, and are usually located at its surface. The prediction of loop structures is a particularly difficult task since they are much less regular and much more variable than ?-helices and ?-sheets and insertions and deletions are most likely to occur in these regions. On the other hand, information about loop structures is often important in that, in many cases, loops have an important functional role: thanks to

their surface location, loops are often involved in interactions with other proteins or in the catalytic mechanism of enzymes and, in some cases, they constitute the nucleation site for protein folding.

The prediction of loops can be based on the sequence patterns in the loop itself, on data base searching methods or on *ab initio* calculations.

Methods based on sequence patterns are generally effective for short loops (3 to 4 residues), especially those connecting adjacent strands of anti-parallel ? sheets, and are based on an exhaustive classification of loops followed by a tabulation of the corresponding specific sequence patterns.

A turn is generally defined as a loop that allows the polypeptide chain to change its direction by 180°. In his early work, Venkatachalam (Venkatachalam 1968; Venkatachalam and Ramachandran 1969) defined a turn as being characterised by the formation of a hydrogen bond between the main chain carboxylic oxygen of the first residue and the amide proton of the third, and identified three conformations (called I, II and III) according to the main chain dihedral angles of residues in the turn (Fig. 5.2a). The mirror images of these turns (denoted I', II', III') are also possible but disfavoured. Beta turns connecting adjacent strands of an antiparallel ? sheet are called ? hair-pin and have been carefully analysed by a number of authors (Chou and Fasman 1977; Ananthanarayanan, et al. 1984; Hollosi, et al. 1985; Sibanda and Thornton 1985; Milner-White and Poet 1986; Wilmot and Thornton 1988; Sibanda, et al. 1989; Sibanda and

Thornton 1991). Two-residue ? hairpins are often found in type I' and II' conformations, rather than I and II as happens for most non-hairpin turns. These conformations are probably preferred for ? hairpins because they allow the correct twisting of the two adjacent ? strands.

Gamma turns are characterised by the presence of a hydrogen bond between the carbonyl group of one residue and the amino group of the amino acid two residues ahead in the sequence and are further classified as classic and inverse, which are the mirror image of each other (Milner-White, et al. 1988; Milner-White 1990) (Fig. 5.2b). Inverse gamma turns are much more frequent than classic ones.

Omega loops have been described and classified (Leszczynski and Rose 1986; Fetrow 1995) as irregular segments of chain where sequentially distant N and C-terminal residues are spatially close (Fig. 5.2c).

Short loops have to introduce a sharp change in the direction of the polypeptide chain and this implies restrictions on the dihedral angles of their residues which can in turn be correlated to the presence of special amino acids such as glycines and prolines. For example, an analysis of the sequence patterns of ? hair-pins in proteins (Wilmot and Thornton 1988) has led to the widely accepted view that type I turns and type II turns are often associated with the presence of a glycine as the third or fourth amino acid of the loop, respectively. Similarly, sequence patterns have been associated with the various turn types and with omega and gamma turns.

The usage of known sequence patterns in comparative modelling, however, would require that the precise length and type of the loop are known a priori and this is rarely the case. Moreover, reconciling the results of different authors on their sequence / structure analysis is not trivial. One of the effects of the intrinsic irregularity of turns and loops is that their definition and nomenclature varies quite substantially. In different instances, the classification is based on their hydrogen bond pattern or on their main chain dihedral angles or on the distance between their end points (Chou and Fasman 1977; Rose, et al. 1983; Sibanda and Thornton 1985; Milner-White, et al. 1988; Wilmot and Thornton 1988; Milner-White 1990; Wilmot and Thornton 1990).

Another problem of all the methods based on sequence patterns is that they rely on the assumption that the structure of the loop is determined by local rather than tertiary interactions. This is not necessarily true and there are cases where it is clear that tertiary interactions can completely overrule the sequence pattern of the loop (Tramontano, et al. 1990).

Medium-sized loops are even more difficult to predict, since the many possible combinations of their main chain dihedral angles make their classification less rigorous.

In general they can be classified according to the type of interactions that stabilise them (Tramontano, et al. 1989). For loops that form compact substructures, the major conformation determinant is the formation of hydrogen bonds to main chain atoms of the loop. For loops having more

extended conformations, the required stabilisation is obtained by packing an inward pointing hydrophobic side chain of the loop between the secondary structure elements connected by the loop. However, it appears that the loop dictates the interactions required to stabilise it, but in different proteins a variety of different topologies can be used to provide these interactions. This implies that this type of classification is not useful in deriving predictive rules.

Data base searching is the default technique used in many model building experiments, especially for medium sized loops, and is based on the observation that segments of similar conformation occur in both related and unrelated proteins (Jones and Thirup 1986).

Methods to predict loops by data base searching assume that some information about the structure of the loop is contained in the regions surrounding it, so that the latter can be used to identify the former. So far, only the information contained in the regions adjacent to the loop in the primary structure of the protein have been taken into account in this approach. This technique, based on the early work of Jones and Thirup (1986) (Jones and Thirup 1986) is generally included as a tool in a number of commercial and academic modelling packages (Dayringer, et al. 1986; Jones and Thirup 1986; Jones, et al. 1990; Vriend 1990).

The basic method consists in searching the data base of solved protein structures for two regions closely matching the segments preceding and following the loop to be modelled ("stems") which are separated by the

same number of residues as those forming the loop to be modelled (Fig. 5.3). The assumption is that the structure of the loop is correlated to the structure of its "stem" regions. In order for this technique to be useful, three things have to be demonstrated. The first is that the loop to model exists in the data base, the second that similar loops have similar stems and thirdly that an equivalent geometric relationship exists between the stems and the loops in the modelled and template structures.

The first assumption has been shown to be true (Jones and Thirup 1986; Tramontano and Lesk 1992), while the other two hypotheses have been shown not to be generally correct in a simulated model building experiment.

The basic steps of the experiment (Tramontano and Lesk 1992) consisted of:

1) searching the data base for loops similar in conformation to specific ones (immunoglobulin hypervariable loops were used for the test), in order to show that the loop could indeed be predicted by using a data base search. The result of this search showed that, with rare exceptions, it was possible to find structurally similar regions in the data base of known structures for all the selected loops, both among immunoglobulin structures and unrelated proteins,

2) searching the data base for regions matching the stems of each loop separated by the appropriate number of residues,

3) comparing the loops selected in step 2 with those selected in step 1.

Such a simulation has a better chance of being successful than a real modelling experiment because, in the test, the stem structures are experimentally determined and not predicted, as would be the case in a model building experiment. Nevertheless, while in some cases a good fit of the stems corresponded to a good fit of the intervening regions, in most cases a good fit of the stem did not imply a good fit of the loop or vice versa. There were also cases where, although both the fit of the stems and the loop were good, the geometric relationship between the two was different.

These results indicate that, although in most cases loops of the desired conformation exist in non-homologous proteins, the information contained in the structure of the adjacent regions is not sufficient to identify them. Consistently with this observation, loop selection by data base searching seems to be able to provide a reasonable result only when homologous protein fragments are used, while the use of nonredundant fragment libraries remains problematic (Bates, et al. 1997).

Other methods for loop structure prediction include molecular simulations, combinatorial searches and subsequent evaluation of loop candidates using conformational energy estimation and a combination of this latter method with data base searching techniques (Rose, et al. 1985; Moult and James 1986; Bruccoleri, et al. 1988; Martin, et al. 1989).

Our understanding of the inter atomic-interactions is neither complete nor exact so that there would be serious computational difficulties in

obtaining the correct evaluation of the energy of the protein's fragments, even if the rest of the model were exact. In a model building experiment, however, the part of the structure into which one is trying to build the loop is affected by an error, no matter what method has been used to predict it. This might seriously affect the result of a detailed free energy calculation.

In other words, on one side the available forcefields (see below) are not sufficiently exact or complete to correctly evaluate the energy of different conformations. On the other, the end points of the loop regions are modelled and are consequently affected by an error which could strongly influence the results of a detailed energy calculation. Furthermore, it has been shown that pentapeptides of identical sequences can have completely different structures in different proteins (Kabsch & Sander, 1984) so that any effective method to predict the folding of a segment of a protein needs to take into account its interaction with the rest of the structure in the calculations.

Research in this field is extremely active, especially since it has been shown that incorrect loop predictions are the major cause of errors in models, but so far the conclusion is that available methods to predict loop conformations are able to identify the correct conformation only in a limited number of cases.

From a practical point of view, however, it is advisable to critically evaluate whether it is really necessary to model the loops: while some of

them are critical for protein functions, others are far from the regions of main interest of the model (e.g., binding or catalytic sites) and their prediction can therefore be omitted.

II.e. Side chain modelling

As the target and template proteins are assumed to be related, the conformation of the side-chains of the conserved residues should be modelled on that of the corresponding residues of template. A number of techniques are usually employed to build the conformations of the mutated ones.

Some are based on 'rotamer libraries' that is a collection of the limited number of conformations that each amino acid side-chain preferentially assumes (Ponder and Richards 1987; Dunbrack and Karplus 1993; Schrauber, et al. 1993; Bower, et al. 1997; Ogata and Umeyama 1997) (Fig. 5.4).

These libraries can differ in the grouping of amino acid used to calculate the statistics of the rotamer distribution, for example by taking into account the local environment of a residue or its backbone angles. In some cases, these libraries, combined with some method to exclude rotamers producing unfavourable steric interactions (Keller, et al. 1995; Bates, et al. 1997), are used to build all the side-chains of the model.

Another commonly used method to build the side chains of non conserved residues is to import the conformational angles of the template

up to where the relative length of the two side-chains permits, using rotamer libraries for the remaining part of the chain (Tramontano 1995).

Energy-based procedures are also used. Usually these procedures start their refinement from a model having the most common rotamers at every position (Rooman, et al. 1991; Thornton 1991; Eisenmenger, et al. 1993; Laughton 1994; Melo and Feytmans 1997).

While methods for modelling side-chain conformation seem to perform rather well when given experimental co-ordinates for the backbone atoms, their accuracy is much lower for protein models and decreases rapidly as the r.m.s.d. between the model and the real structure increases (Martin, et al. 1997). This might indicate that an improvement in this area could be automatically achieved as a consequence of improvements in backbone modelling.

III. Model refinement

The steps described above usually lead to a (fairly) complete model.

The next stage of the process consists in inspecting it, both visually and through the use of specific programs, to evaluate and optimise it.

In principle, incorrect stereochemistry, unfavourable steric interactions and poorly packed regions have to be fixed. This can be done through manual interventions followed by few cycles of energy minimisation or geometric refinement. It is important to emphasise that neither method can substantially modify the starting model and consequently cannot correct large errors.

Many attempts have been made to obtain a global refinement of a protein model using energy minimisation techniques or molecular dynamics. In blind tests, energy minimisation and molecular dynamics did not improve the models and often models built without any further refinement were closer to the real structures than those 'optimised' using various combinations of these methods (Martin, et al. 1997).

Extensive energetic refinement, with whatever method, besides failing to improve the model, also has a major disadvantage. Regions of the model that are less reasonable from a structural point of view, are likely to be those containing errors, due to either the initial alignment or the model building procedure used. Their optimisation, often obtained by moving atoms in the neighbouring regions, will hide this effect and will make every part of the model equally 'good'. For the model to be used correctly, it is important to be able to highlight 'problematic' areas of the structures that should be used with much more caution.

IV. Expected accuracy of an homology model

How reliable is a model? As we said before, the overall quality of models is highly dependent on the degree of similarity of the target with the parent structure and on the quality of the sequence alignment (Chothia and Lesk 1986; Hilbert, et al. 1993; Mosimann, et al. 1995; Martin, et al. 1997). Both factors are related to the degree of sequence similarity and to the number of insertions and deletions between target and parent.

In principle the relationship between sequence similarity and structural divergence could be used to evaluate a priori the accuracy of a model. However this is based on the assumption that the alignment is correct and that no improvement has been introduced by optimisation methods.

It is therefore important to test the quality of models in blind tests and the field of protein structure prediction has received a new impulse from the CASP experiments.

The impact of blind assessment experiments has been high in the field of comparative modelling. The technique has been used for so many years and in so many instances that it was taken for granted that serious progress in all its aspects was ongoing and that very good models could be built, even by using just automatic servers or programs.

This is true for models of the core regions of proteins when the template is quite similar in sequence, but the results of the blind tests were far from exciting for medium to low homology models.

In the first CASP experiment, no correct prediction of loops was obtained, the accuracy of side chain positioning was surprisingly low, the alignments contained many errors and some of the models were even stereochemically wrong (Mosimann, et al. 1995).

Some progress was achieved in CASP2, where predictors were challenged to predict 9 targets with a sequence identity to the best template ranging between 20% and 85% (Martin, et al. 1997).

One lesson from CASP1 was that automatic alignments should be refined and optimised manually and almost all groups did so in CASP2. One interesting example of how well this can work is illustrated below by the case of T28 (endoglucanase I).

An automatic alignment in this case provides:

```
TARGET:NGSPSGNLVSITRKYQQNGVDIPSAQPGGDTISSCPS------ASAY---GGL

PARENT:SGAINRYYVQNGVTFQQPNAELGSYSGNELNDDYCTAEEAEFGGSSFSDKGGL
```

while the correct alignment turned out to be:

```
TARGET:NGSPSGNLVSITRKYQQNGVDIPS-AQ------PG-GDTISSCP----------SASAYGGL
PARENT:S------G-AINRYYVQNGVTFQ-QPNAELGSYSGNELNDDYCTAEEAEFGGSSF-SDKGGL
```

and the manually edited alignment by Samudrala et al. (Samudrala and Moult 1997):

```
TARGET:NGSPSGNLVSITRKYQQNGVDIPSA-------QPGGDTISSCP---------SASAYGGL
PARENT:-------SGAINRYYVQNGVTFQQPNAELGSYSGNELNDDYCTAEEAEFGGSSFSDKGGL
```

The improvement is substantial in this case since the number of correctly aligned residues is 29 in the manually edited alignment and only 1 in the automatic one!

It also became clear that the possibility of achieving a correct alignment is not only dependent on the sequence identity between target and templates, but also on the protein itself. In other words, while a high sequence identity almost guarantees a reasonable alignment, there are cases where similar sequence identity values (20% and 26%) produce alignments with a mean shift of 10 and 0.2 residues, respectively (Martin, et al. 1997). This specific example is for proteins pectate lyase and

stellacyanin and seems to reflect both the type of protein (pectate lyase has a repeated structure and this might confuse alignment programs) and the number and distributions of insertion and deletions.

In CASP2 some loops were finally predicted moderately well, and there was improvement in modelling side chains.

The open issues highlighted by the assessment of comparative modelling techniques are however not very different from those of the first CASP:

Alignment accuracy is still the major factor in determining the quality of a model, and its correctness below 25% sequence identity is still very low, the deviation between modelled and experimental loops is large, and it the main factor in the high values of total r.m.s.d., correct inheritance from the parent (in terms of distinguishing which parts can be copied into the model and which cannot) is the most important factor. The best models are those which deviate less from the parent structure. In other words, any attempt to model *de novo* regions of the protein or more sophisticated approaches which inherit their structures less directly from the parents seem to perform less well.

It is important to note, though, that the most conserved regions in proteins are those that have an important structural and/or functional role and these regions are often those modelled with higher accuracy; therefore, in spite of their possible shortcomings, models built by homology generally contain a wealth of practically useful information and are often instrumental in interpreting experimental data, in planning new

experiments and in guiding the design of modified proteins (Sollazzo, et al. 1990; Savino, et al. 1993; Orlandini, et al. 1994; Pizzi, et al. 1994; Scarborough and Dunn 1994; Amati, et al. 1995; Ammendola, et al. 1995; Helmer-Citterich, et al. 1995; De Francesco, et al. 1996; Failla, et al. 1996; Luo, et al. 1996; Jackson, et al. 1997; Komissarov, et al. 1997; Neddermann, et al. 1997; Starling, et al. 1997).

V. Practical remarks

Each step of the procedure to build homology models introduces errors because of the underlying theoretical issues, but there are also technical points that may seem more trivial, but are nevertheless equally responsible for errors and frustrations.

The problems related to understanding and analysing the result of the data base search that needs to be performed to select the template have been discussed before.

However, once the template has been selected, and a sequence alignment has been obtained, the alignment should be optimised using a number of criteria to take into account the available information about the template protein structure, the structural superposition of homologous structures and the multiple sequence alignment of the family of the target protein.

The structural alignment of two protein structures is in itself a complex problem, and involves the selection of a number of parameters.

In principle one would like to obtain the best structural superposition of two structures within a given r.m.s.d. threshold, but the problem does not

necessarily have a unique answer. Methods for structural superposition are becoming more clever (Taylor and Orengo 1989a; Taylor and Orengo 1989b; Orengo and Taylor 1990; Rose and Eisenmenger 1991; Holm and Sander 1992; Orengo, et al. 1992; Holm and Sander 1993; Orengo, et al. 1993; Orengo and Taylor 1993; Holm and Sander 1994; May and Johnson 1994; Taylor, et al. 1994; Holm and Sander 1995b; Holm and Sander 1995a; Maiorov and Crippen 1995; May and Johnson 1995; Falicov and Cohen 1996; Holm and Sander 1996a; Holm and Sander 1996b; Michie, et al. 1996; Orengo and Taylor 1996; Holm and Sander 1997; Orengo, et al. 1997; Singh and Brutlag 1997), but the user still has to decide what 'structurally conserved' means.

A possible procedure is to:

1. obtain a structural superposition of the two proteins,
2. identify those regions that have the same secondary structure,
3. least square fit of these zones,
4. extend the selected zones at each end adding a residue at the time while the distance between equivalent C-? is less than a threshold, for example 3.0 Å.

This procedure is not easy to apply to more than two structures, and might produce somewhat different results according to which structural superposition method is used in step 1, which definition of secondary structure is used in step 2 and which threshold is selected in step 4, but

can be considered as a useful starting point to analyse structural conservation.

As far as the multiple sequence alignment of the target protein family is concerned, there are a number of sequence alignment editing tools which are quite useful. One practical suggestion is to use colour codes for each amino acid (a feature now provided by a number of programs) because this is usually very useful in detecting regions whose alignment can be manually improved.

A number of parameters also have to be selected when modelling insertions and deletions by data base searching. The alignment will show a number of residues to be inserted/deleted, but it is advisable to look in the data base for a region including at least one residue on each side of the insertion/deletion. In other words in a case such as

```
template ADFRLADGTRCFRGT
target   LEFTLVD-SKCWKAS
         |stem|   |stem|
```

the most advisable thing would be to search for regions similar to the stems and separated by 2 residues. Obviously the extent of the stems also has to be selected. Usual default values in modelling packages are about 5 residues, however this parameter has to be selected on a case by case basis also taking into account the secondary structure of the stems. For example, if the preceding stem is in an ? helix, at least 45 residues should be used to guarantee that the matching regions would also be an ? helix, while for a ? strand a shorter region might be sufficient. In any

case, it is always advisable to perform the search varying the dimensions of the stems to verify how much this affects the type of structures selected in the data base.

Usually a number of regions are proposed as appropriate by modelling packages and it is not straightforward to decide which one to choose. Once again a number of rules of thumb can be given:

1. The selected region should not be bumping into the backbone of the conserved regions of the protein or with C? atoms of surrounding residues.
2. If the proposed loops have different conformations, the region which has matching pattern of glycines and prolines should be selected.
3. Among similar structures, the one with a lower r.m.s.d. of the stems should be selected.

After this step is completed, a modeller will encounter what may seem a minor problem (but is not considered so by whoever builds a model manually or by the authors of automatic modelling programs) and that is the handling of the Protein Data Bank entries and of their residue numbering.

Records in a pdb file (Bernstein, et al. 1977) look like:

```
ATOM   49     N     ILE   Z   16      33.126  59.594  12.174  0.00   0.00     1      3TPI 155
ATOM   50     CA    ILE   Z   16      31.81   59.24   12.722  0.00   0.00     1      3TPI 156
ATOM   51     C     ILE   Z   16      30.843  60.42   12.748  0.00   0.00     1      3TPI 157
ATOM   52     O     ILE   Z   16      31.053  61.388  13.523  0.00   0.00     1      3TPI 158
ATOM   53     CB    ILE   Z   16      31.965  58.688  14.158  0.00   0.00     1      3TPI 159
ATOM   54     CG1   ILE   Z   16      32.941  57.499  14.229  0.00   0.00     1      3TPI 160
ATOM   55     CG2   ILE   Z   16      30.612  58.32   14.798  0.00   0.00     1      3TPI 161
ATOM   56     CD1   ILE   Z   16      32.349  56.23   13.587  0.00   0.00     1      3TPI 162
```

```
ATOM   57     N     VAL   Z   17     29.731  60.251  12.065  0.00   0.00     1      3TPI 163
ATOM   58     CA    VAL   Z   17     28.581  61.149  12.183  0.00   0.00     1      3TPI 164
ATOM   59     C     VAL   Z   17     27.533  60.535  13.105  0.00   0.00     1      3TPI 165
ATOM   60     O     VAL   Z   17     27.577  59.301  13.343  0.00   0.00     1      3TPI 166
ATOM   61     CB    VAL   Z   17     27.971  61.439  10.798  0.00   0.00     1      3TPI 167
ATOM   62     CG1   VAL   Z   17     28.965  62.183  9.886   0.00   0.00     1      3TPI 168
ATOM   63     CG2   VAL   Z   17     27.427  60.167  10.12   0.00   0.00     1      3TPI 169
....
ATOM   1233   N     CYS   Z   182    19.441  67.224  29.58   1.00   14.82           3TPI1339
ATOM   1234   CA    CYS   Z   182    20.456  67.33   28.488  1.00   14.82           3TPI1340
ATOM   1235   C     CYS   Z   182    20.739  65.925  28.089  1.00   14.82           3TPI1341
ATOM   1236   O     CYS   Z   182    20.71   65.053  28.975  1.00   14.82           3TPI1342
ATOM   1237   CB    CYS   Z   182    21.785  67.927  28.982  1.00   14.82           3TPI1343
ATOM   1238   SG    CYS   Z   182    21.809  69.712  29.071  1.00   23.37           3TPI1344
ATOM   1239   N     ALA   Z   183    20.936  65.758  26.824  1.00   19.04           3TPI1345
ATOM   1240   CA    ALA   Z   183    21.297  64.486  26.279  1.00   19.04           3TPI1346
ATOM   1241   C     ALA   Z   183    22.26   64.676  25.102  1.00   19.04           3TPI1347
ATOM   1242   O     ALA   Z   183    22.216  65.695  24.403  1.00   19.04           3TPI1348
ATOM   1243   CB    ALA   Z   183    20.071  63.625  25.866  1.00   19.04           3TPI1349
ATOM   1244   N     GLY   Z   184A   23.147  63.766  24.97   1.00   24.01           3TPI1350
ATOM   1245   CA    GLY   Z   184A   24.18   63.798  23.932  1.00   24.01           3TPI1351
ATOM   1246   C     GLY   Z   184A   25.595  63.678  24.5    1.00   24.01           3TPI1352
ATOM   1247   O     GLY   Z   184A   25.854  62.902  25.456  1.00   24.01           3TPI1353

       Atom   Atom  Res.  c  Res.    X       Y       Z       Occ.   B-fact  Note
       number name  name  h  number
                          a
                          i
                          n
```

The reader should notice that residue numbers of entries in the PDB data base are not numbers but strings, containing characters and numbers, not necessarily consecutive and not necessarily starting from 1.

Residue numbers can be followed by letters (184A). Very often this reflects the existence of a common residue numbering scheme for the protein family. The example reported here is in fact that of trypsin and serine proteases usually follow what is called the 'chymotrypsin numbering scheme'. Since the PDB entry 3tpi (Huber, et al. 1974) contains both trypsin and its proteic inhibitor, the enzyme is identified by a chain (Z in this case).

Also note in this example that residues 16 and 17 have occupancy 0, that is they cannot be seen in the structure. A footnote in the file in fact explains that:

```
AN OCCUPANCY OF 0.0 INDICATES THAT NO SIGNIFICANT DENSITY WAS
FOUND IN THE FINAL FOURIER MAP.
```

In the same file there are notes such as:

```
THERE IS NO SIGNIFICANT ELECTRON DENSITY IN THE FINAL
FOURIER MAP FOR THE N-TERMINUS OF THE ZYMOGEN FROM VAL Z 10
THROUGH GLY Z 18 AND THIS DATA ENTRY CONTAINS NO
COORDINATES FOR VAL Z 10 THROUGH LYS Z 15.
```

which indicates the importance of reading the remark records of the entry.

In general, modelling programs will preserve the original residue numbering of any loop selected in a data base search, will not verify whether the selected region has all atoms with full occupancy and will not record the original occupancy values. After a few insertions and deletions, the issue of finding the regions corresponding to a given part of the alignment in the structure will become quite annoying and there will be no way to recover the original occupancy values or any of the authors' remarks for the regions used to construct the loops. This implies that the original entry of each of the PDB files used in a modelling experiment should be manually checked, in particular with respect to occupancy and remarks of the authors.

Another issue that can arise both during the selection of the template and of the putative loops is that of having to use NMR structures. NMR

produces an ensemble of structures, a number of which are usually included in the PDB entry together with an 'average structure'. Care should be taken in understanding which structure is being used and what is the extent of variation of the structure in the different models of the molecule.

Figure legends

Fig 5.1 Number of protein structures deposited each year into the Brookhaven Protein Data Bank (PDB) (Bernstein, et al. 1977).

Fig. 5.2 Types of turns in protein structures. a) Beta turns. Type I and I' are shown on the top, Type II and II' in the middle and Type III and III' on the bottom. b) Gamma turn and Inverse gamma turn. c) An example of an omega turn.

Fig. 5.3 Example of loop prediction by data base searching. Left: backbone of a putative template structure with a 6 residue long loop. Right: C-alpha trace of the template loop and of two loops of length 3 selected because their stems (5 residues before and 5 residues after the loop) are very similar to the stems of the template structure.

Fig 5.4 The four most common rotamers of tyrosine. They are shown after optimal superposition of the backbone.

References

Altschul, S. (1989) Gap costs for multiple sequence alignment. *J Theor Biol* **138**:297-309.

Amati, V., Werge, T., Cattaneo, A.& Tramontano, A. (1995) Identifying a putative common binding site shared by substance P receptor and an anti-substance P monoclonal antibody. *Protein Eng* **8**:403-408.

Ammendola, S., Raucci, G., Incani, O.*et al.* (1995) Replacing the glutamate ligand in the structural zinc site of Sulfolobus solfataricus alcohol dehydrogenase with a cysteine decreases thermostability. *Protein Eng* **8**:31-37.

Ananthanarayanan, V., Brahmachari, S.& Pattabiraman, N. (1984) Proline-containing beta-turns in peptides and proteins: analysis of structural data on globular proteins. *Arch Biochem Biophys* **232**:482-495.

Bates, P., Jackson, R.& Sternberg, M. (1997) Model building by comparison: a combination of expert knowledge and computer automation. *Proteins* **Suppl 1**:59-67.

Bernstein, F., Koetzle, T., Williams, G.*et al.* (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem* **80**:319-324.

Bower, M., Cohen, F.& Dunbrack RL, J. (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol* **267**:1268-1282.

Bruccoleri, R., Haber, E.& Novotny, J. (1988) Structure of antibody hypervariable loops reproduced by a conformational search algorithm [published erratum appears in Nature 1988 Nov 17;;336(6196):266]. *Nature* **335**:564-568.

Chothia, C.& Lesk, A. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**:823-826.

Chou, P.& Fasman, G. (1977) Beta-turns in proteins. *J Mol Biol* **115**:135-175.

Crippen, G. (1977) Correlation of sequence and tertiary structure in globular proteins. *Biopolymers* **16**:2189-2201.

Dayringer, H.E., Tramontano, A., Sprang, S.R.& Fletterick, R.J. (1986) Interactive program for visualization and modeling of protein, nucleic acids and small molecules. *J. Mol. Graphics* **4**:82-87.

De Francesco, R., Urbani, A., Nardi, M.*et al.* (1996) A zinc binding site in viral serine proteinases. *Biochemistry* **35**:13282-13287.

Dunbrack, R., Jr& Karplus, M. (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* **230**:543-574.

Eisenmenger, F., Argos, P.& Abagyan, R. (1993) A method to configure protein side-chains from the main-chain trace in homology modelling. *J Mol Biol* **231**:849-860.

Failla, C., Pizzi, E., De Francesco, R.& Tramontano, A. (1996) Redesigning the substrate specificity of the hepatitis C virus NS3 protease. *Fold Des* **1**:35-42.

Falicov, A.& Cohen, F. (1996) A surface of minimum area metric for the structural comparison of proteins. *J Mol Biol* **258**:871-892.

Fetrow, J. (1995) Omega loops: nonregular secondary structures significant in protein function and stability. *FASEB J* **9**:708-717.

Gotoh, O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol* **264**:823-838.

Gracy, J., Chiche, L.& Sallantin, J. (1993) Improved alignment of weakly homologous protein sequences using structural information. *Protein Eng* **6**:821-829.

Grindley, H., Artymiuk, P., Rice, D.& Willett, P. (1993) Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol* **229**:707-721.

Helmer-Citterich, M., Rovida, E., Luzzago, A.& Tramontano, A. (1995) Modelling antibody-antigen interactions: ferritin as a case study. *Mol Immunol* **32**:1001-1010.

Henneke, C. (1989) A multiple sequence alignment algorithm for homologous proteins using secondary structure information and optionally keying alignments to functionally important sites. *Comput Appl Biosci* **5**:141-150.

Hilbert, M., Bohm, G.& Jaenicke, R. (1993) Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins* **17**:138-151.

Hollosi, M., Kawai, M.& Fasman, G. (1985) Studies on proline-containing tetrapeptide models of beta-turns. *Biopolymers* **24**:211-242.

Holm, L.& Sander, C. (1992) Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins* **14**:213-223.

Holm, L.& Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol* **233**:123-138.

Holm, L.& Sander, C. (1994) Parser for protein folding units. *Proteins* **19**:256-268.

Holm, L.& Sander, C. (1995a) 3-D lookup: fast protein structure database searches at 90% reliability. *Ismb* **3**:179-187.

Holm, L.& Sander, C. (1995b) Dali: a network tool for protein structure comparison. *Trends Biochem Sci* **20**:478-480.

Holm, L.& Sander, C. (1996a) Alignment of three-dimensional protein structures: network server for database searching. *Methods Enzymol* **266**:653-662.

Holm, L.& Sander, C. (1996b) The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res* **24**:206-209.

Holm, L.& Sander, C. (1997) Decision support system for the evolutionary classification of protein structures. *Ismb* **5**:140-146.

Huber, R., Kukla, D., Bode, W.*et al.* (1974) Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. II. Crystallographic refinement at 1.9 A resolution. *J Mol Biol* **89**:73-101.

Jackson, T., Cooper, S.& Church, F. (1997) Assessment of the interaction between urokinase and reactive site mutants of protein C inhibitor. *J Protein Chem* **16**:819-828.

Jones, T.& Thirup, S. (1986) Using known substructures in protein model building and crystallography. *EMBO J* **5**:819-822.

Jones, T.A., Bergdoll, M.& Kjeldgaard, M. (1990) . Springer Verlag, New York.

Kanaoka, M., Kishimoto, F., Ueki, Y.& Umeyama, H. (1989) Alignment of protein sequences using the hydrophobic core scores. *Protein Eng* **2**:347-351.

Keller, D., Shibata, M., Marcus, E., Ornstein, R.& Rein, R. (1995) Finding the global minimum: a fuzzy end elimination implementation. *Protein Eng* **8**:893-904.

Koch, I.& Lengauer, T. (1997) Detection of distant structural similarities in a set of proteins using a fast graph-based method. *Ismb* **5**:167-178.

Koch, I., Lengauer, T.& Wanke, E. (1996) An algorithm for finding maximal common subtopologies in a set of protein structures. *J Comput Biol* **3**:289-306.

Komissarov, A., Marchbank, M., Calcutt, M., Quinn, T.& Deutscher, S. (1997) Site-specific mutagenesis of a recombinant anti-single-stranded DNA Fab. Role of heavy chain complementarity-determining region 3 residues in antigen interaction. *J Biol Chem* **272**:26864-26870.

Laughton, C. (1994) A study of simulated annealing protocols for use with molecular dynamics in protein structure prediction. *Protein Eng* **7**:235-241.

Leszczynski, J.& Rose, G. (1986) Loops in globular proteins: a novel category of secondary structure. *Science* **234**:849-855.

Luo, Y., Gabriel, J., Wang, F.*et al.* (1996) Molecular modeling and deletion mutagenesis implicate the nuclear translocation sequence in structural integrity of fibroblast growth factor-1. *J Biol Chem* **271**:26876-26883.

Luo, Y., Lai, L., Xu, X.& Tang, Y. (1993) Defining topological equivalences in protein structures by means of a dynamic programming algorithm. *Protein Eng* **6**:373-376.

Maiorov, V.& Crippen, G. (1995) Size-independent comparison of protein three-dimensional structures. *Proteins* **22**:273-283.

Mandal, C.& Linthicum, D. (1993) PROGEN: an automated modelling algorithm for the generation of complete protein structures from the alpha-carbon atomic coordinates. *J Comput Aided Mol Des* **7**:199-224.

Martin, A., Cheetham, J.& Rees, A. (1989) Modeling antibody hypervariable loops: a combined algorithm. *Proc Natl Acad Sci U S A* **86**:9268-9272.

Martin, A., MacArthur, M.& Thornton, J. (1997) Assessment of comparative modeling in CASP2. *Proteins* **Suppl 1**:14-28.

May, A.& Blundell, T. (1994) Automated comparative modelling of protein structures. *Curr Opin Biotechnol* **5**:355-360.

May, A.& Johnson, M. (1994) Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization. *Protein Eng* **7**:475-485.

May, A.& Johnson, M. (1995) Improved genetic algorithm-based protein structure comparisons: pairwise and multiple superpositions. *Protein Eng* **8**:873-882.

Melo, F.& Feytmans, E. (1997) Novel knowledge-based mean force potential at atomic level. *J Mol Biol* **267**:207-222.

Michie, A., Orengo, C.& Thornton, J. (1996) Analysis of domain structural class using an automated class assignment protocol. *J Mol Biol* **262**:168-185.

Milner-White, E. (1990) Situations of gamma-turns in proteins. Their relation to alpha-helices, beta-sheets and ligand binding sites. *J Mol Biol* **216**:386-397.

Milner-White, E.& Poet, R. (1986) Four classes of beta-hairpins in proteins. *Biochem J* **240**:289-292.

Milner-White, E., Ross, B., Ismail, R., Belhadj-Mostefa, K.& Poet, R. (1988) One type of gamma-turn, rather than the other gives rise to chain-reversal in proteins. *J Mol Biol* **204**:777-782.

Mosimann, S., Meleshko, R.& James, M. (1995) A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins* **23**:301-317.

Moult, J.& James, M. (1986) An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins* **1**:146-163.

Neddermann, P., Tomei, L., Steinkuhler, C.*et al.* (1997) The nonstructural proteins of the hepatitis C virus: structure and functions. *Biol Chem* **378**:469-476.

Ogata, K.& Umeyama, H. (1997) Prediction of protein side-chain conformations by principal component analysis for fixed main-chain atoms. *Protein Eng* **10**:353-359.

Orengo, C., Brown, N.& Taylor, W. (1992) Fast structure alignment for protein databank searching. *Proteins* **14**:139-167.

Orengo, C., Flores, T., Taylor, W.& Thornton, J. (1993) Identification and classification of protein fold families. *Protein Eng* **6**:485-500.

Orengo, C., Michie, A., Jones, S.*et al.* (1997) CATH--a hierarchic classification of protein domain structures. *Structure* **5**:1093-1108.

Orengo, C.& Taylor, W. (1990) A rapid method of protein structure-alignment. *J Theor Biol* **147**:517-551.

Orengo, C.& Taylor, W. (1993) A local alignment method for protein structure motifs. *J Mol Biol* **233**:488-497.

Orengo, C.& Taylor, W. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol* **266**:617-635.

Orlandini, M., Santucci, A., Tramontano, A., Neri, P.& Oliviero, S. (1994) Cloning, characterization, and modeling of a monoclonal anti-human

transferrin antibody that competes with the transferrin receptor. *Protein Sci* **3**:1476-1484.

Pauling, l., Corey, R.B.& Branson, H.R. (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci USA* **37**:205-211.

Peitsch, M. (1996) ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem Soc Trans* **24**:274-279.

Peitsch, M. (1997) Large scale protein modelling and model repository. *Ismb* **5**:234-236.

Peitsch, M., Herzyk, P., Wells, T.& Hubbard, R. (1996) Automated modelling of the transmembrane region of G-protein coupled receptor by Swiss-model. *Receptors Channels* **4**:161-164.

Pizzi, E., Tramontano, A., Tomei, L.*et al.* (1994) Molecular model of the specificity pocket of the hepatitis C virus protease: implications for substrate recognition. *Proc Natl Acad Sci U S A* **91**:888-892.

Ponder, J.& Richards, F. (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* **193**:775-791.

Rippmann, F.& Taylor, W. (1991) Visualization of structural similarity in proteins. *J Mol Graph* **9**:169-174, 163-164.

Rooman, M., Kocher, J.& Wodak, S. (1991) Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions. *J Mol Biol* **221**:961-979.

Rose, G., Gierasch, L.& Smith, J. (1985) Turns in peptides and proteins. *Adv Protein Chem* **37**:1-109.

Rose, G., Young, W.& Gierasch, L. (1983) Interior turns in globular proteins. *Nature* **304**:654-657.

Rose, J.& Eisenmenger, F. (1991) A fast unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm. *J Mol Evol* **32**:340-354.

Rufino, S.& Blundell, T. (1994) Structure-based identification and clustering of protein families and superfamilies. *J Comput Aided Mol Des* **8**:5-27.

Russell, R.& Barton, G. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* **14**:309-323.

Sali, A. (1995) Modeling mutations and homologous proteins. *Curr Opin Biotechnol* **6**:437-451.

Sali, A.& Blundell, T. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**:779-815.

Samudrala, R.& Moult, J. (1997) Handling context-sensitivity in protein structures using graph theory: bona fide prediction. *Proteins* **Suppl 1**:43-49.

Savino, R., Lahm, A., Giorgio, M.*et al.* (1993) Saturation mutagenesis of the human interleukin 6 receptor-binding site: implications for its three-dimensional structure. *Proc Natl Acad Sci U S A* **90**:4067-4071.

Scarborough, P.& Dunn, B. (1994) Redesign of the substrate specificity of human cathepsin D: the dominant role of position 287 in the S2 subsite. *Protein Eng* **7**:495-502.

Schmidt, R., Gerstein, M.& Altman, R. (1997) LPFC: an Internet library of protein family core structures. *Protein Sci* **6**:246-248.

Schrauber, H., Eisenhaber, F.& Argos, P. (1993) Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol* **230**:592-612.

Sibanda, B., Blundell, T.& Thornton, J. (1989) Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J Mol Biol* **206**:759-777.

Sibanda, B.& Thornton, J. (1985) Beta-hairpin families in globular proteins. *Nature* **316**:170-174.

Sibanda, B.& Thornton, J. (1991) Conformation of beta hairpins in protein structures: classification and diversity in homologous structures. *Methods Enzymol* **202**:59-82.

Singh, A.& Brutlag, D. (1997) Hierarchical protein structure superposition using both secondary structure and atomic representations. *Ismb* **5**:284-293.

Sollazzo, M., Castiglia, D., Billetta, R., Tramontano, A.& Zanetti, M. (1990) Structural definition by antibody engineering of an idiotypic determinant. *Protein Eng* **3**:531-539.

Venkatachalam, C.& Ramachandran, G. (1969) Conformation of polypeptide chains. *Annu Rev Biochem* **38**:45-82.

Vriend, G. (1990) WHAT IF: a molecular modelling and drug design program. *J. Mol. Graphics* **8**:52-56.

Wilmot, C.& Thornton, J. (1988) Analysis and prediction of the different types of beta-turn in proteins. *J Mol Biol* **203**:221-232.

Wilmot, C.& Thornton, J. (1990) Beta-turns and their distortions: a proposed new nomenclature. *Protein Eng* **3**:479-493.

Zuker, M.& Somorjai, R. (1989) The alignment of protein structures in three dimensions. *Bull Math Biol* **51**:55-78.

IMPLICATIONS FOR PROTEIN DESIGN

As discussed above, the folding code is degenerate, so that many proteins, even when they do not share any significant sequence homology, adopt similar folds. This implies that designing a sequence able to adopt a given fold should in principle be easier than predicting which fold a given sequence will assume.

The *de novo* design of proteins has in fact met with success albeit with some limitations [144, 145, 4, 6, 146, 7, 8, 147]. To refine the process and improve our understanding of the rules that govern protein folding a major effort has also been devoted to the detailed structural characterisation of such designed molecules [5, 148, 149, 9, 150, 12]. However, because our understanding of the protein folding code is still fairly rudimentary the option of recruiting known protein folds as frameworks for the insertion of functional sites, or the modification of existing enzymatic activity, has proved to be more viable and has attracted the most attention [151, 10, 11, 14] (Fig. 2).

The three major challenges that protein design has to face are: i) is the designed sequence compatible with the desired fold? ii) is the selected fold the most favourable for that sequence? iii) does a folding pathway for the fold exist?

The potential function developed for fold recognition and secondary structure prediction methods should be able to help to answer the first question [152]. Fold recognition techniques and folding simulations might become reliable enough to set the other issues so that the new tools developed for protein structure prediction are giving a new impulse also to the field of protein design.

Being able to design proteins for specific functions would have a tremendous impact in many areas of biology and medicine, but would also represent a key step in our understanding of the rules relating sequence to structure in proteins. Native proteins are a

biased sample of the set of possible solutions of the folding problem, since they have been obtained through an enormous number of steps of evolution and selection and since they have constraints imposed by their function and interactions. Designed proteins could allow us to better highlight those properties of protein structure which still escape our understanding.

TRENDS AND FUTURE PERSPECTIVES

It has been suggested and shown in many instances [153, 154] that the combined use of prediction results coming from different methods and of experimental data is one way to improve the quality of the final model of a protein.

A major issue is therefore to be able to compare these data, which are of different type and dimensionality and to verify which model or part of a model is consistent with the larger subset of them.

In our laboratory, we have developed GLASS, a novel tool to address this issue. The system we have implemented is a general platform to read, visualise, project into three-dimensions and compare the results of different structure prediction methods. It also allows to assess the consistency of the model(s) with experimental data and to compare selected parameters calculated for a model with the distribution observed in real protein structures (Fig. 3).

A development version of GLASS was used during the IRBM structure prediction practical workshop [153, 154] on a set of target proteins where it was found to be extremely useful both to compare the results from different prediction methods and to map known experimental data onto the putative models by all the participants.

We feel that this system is equally needed by the users of the many different prediction methods available and by theoreticians who can use it as a workbench to rapidly test new ideas for evaluating the likelihood of different models and is likely that predicting a protein structure with techniques other than homology modelling will become faster and more reliable as more tools for evaluating alternative models and to automatically check the consistency of different results will be added to systems such as GLASS.

The big challenge is whether the automation of both the prediction and evaluation procedures will allow to create a 'prediction database' containing predictions at different levels of detail for any sequence. Given the large number of sequences of unknown structure being generated by the genomic sequencing projects, automation of the prediction and evaluation steps, and the possibility of making the results automatically available via the Internet would provide a valuable resource for theoreticians and experimentalists, and might be a key step in predicting the function of a protein sequence, and understanding its mechanism of action, which is after all, the final goal of protein structural studies.

Acknowledgements

We are grateful to all participants of the 1995 IRBM Workshop 'Frontiers of Protein Structure Prediction' for their help in evaluating GLASS and Tim Hubbard Lahm for many helpful discussions. RL is supported by EEC contract # BIO4-CT96-5034.

References

1. Fuh, G., Cunningham, B.C., Fukunaga, R., Nagata, S., Goeddel, D.V. and Wells, J.A.: Rational design of potent antagonists to human growth hormone receptor. Science 1992; 256:1677-1680.

2. Savino, R., Ciapponi, L., Lahm, A., Demartis, A., Cabibbo, A., Toniatti, C., DelmastroAltamura, S. and Ciliberto, G.: Rational design of a receptor super-antagonist of human interleukin-6. EMBO J. 1994; 13:5863-5870.

3. Savino, R., Lahm, A., Salvati, A.L., Ciapponi, L., Sporeno, E., Altamura, S., Paonessa, G., Toniatti, C. and Ciliberto, G: Generation of interleukin-6 receptor antagonist by molecular-modelling guided mutagenesis of residues important for gp130 activation. EMBO J. 1994; 13:1357-1367.

4. Regan, L. and DeGrado, W.F.: Characterization of a helical protein designed from first principles. Science 1988; 241:976-978.

5. Hubbard, T.J. and Blundell, T.L. In: van Gunsteren WF and Weiner PK (eds.) Computer simulations of Biomolecular systems: Theoretical and experimental applications. Leiden: ESCOM, 1989; 168-182.

6. Richardson, J.S. and Richardson, D.C.: The *de novo* design of protein structures. Trends Biochem. Sci. 1989; 14:304-309.

7. Hecht, M.H., Richardson, J.S., Richardson, D.C. and Odgen, R.C.: *De novo* design, expression, and characterization of felix: a four-helix bundle protein of native-like sequence. Science 1990; 249:884-891.

8. Fedorov, A.N., Dolgikh, D.A., Chemeris, V.V., Chernov, B.K., Finkelstein, A.V., Schulga, A.A., Alakhov, Y.B., Kirpichnikov, M.P. and Ptitsyn, O.B.: *De novo* design, synthesis and study of abebetin a polypeptide with a predetermined three-dimensional structure. Probing the structure at the nanogram level. J. Mol. Biol. 1992; 225:927-931.

9. Sander, C., *et al.*: Protein design on computers. Five new proteins: Shpilka, Grendel, Fingerclasp, Leather, and Aida. Proteins 1992; 12:105-10.

10. Pessi, A., Bianchi, A., Crameri, A., Venturini, S., Tramontano, A. and Sollazzo, M.: A designed metal binding protein with a novel fold. Nature 1993; 362:367-369.

11. Recktenwald, A., Schomburg, D. and Schimd, R.D.: Protein engineering and design: method and industrial relevance. J. Biotechnol. 1993; 28:1-23.

12. Finkelstein, A.V., Hubbard, T.J.P., Lesk, A.M., Moult, J.M., Sander, C., Tramontano, A. and Vriend, G. Protein Design on Computers (ProDes94). Heidelberg, Germany: EMBL, 1994.

13. Rees, A.R., Staunton, D., Webster, D.M., Searcle, S.J., Henry, A.H. and Pedersen, J.T.: Antibody design: beyond the natural limits. Trends Biotechnol. 1994; 12:199-206.

14. Tramontano, A., Bianchi, E., Venturini, S., Martin, F., Pessi, A. and Sollazzo, M.: Making of the Minibody: a designed b-protein for the displaying of conformationally constrained peptides. J. Mol. Recognition 1994; 7:9-24.

15. Saragovi, H.U., Fitzpatrick, D., Raktabutr, A., Nakanishi, H., Kahn, M. and Green, M.J.: Design and synthesis of a mimetic from an antibody complementarity-determining region. Science 1991; 253:792-795.

16. Saragovi, H.U., Green, M.I., Chrusciel, R.A. and Kahn, M.: Loops and secondary structure mimetics: development and applications in basic science and rational drug design. BioTech 1992; 10:773-778.

17. Marshall, G.R., Barry, C.D., Bosshard, H.E., Dammkoheler, R.A. and Dunn, D.A. In: Olson EC and Christofferson RE (eds.) Computer-Assisted Drug Design Washington, DC: American Chemical Society, 1979; 205-226.
18. Colman, P.: Structure-based drug design. Current Opinion in Structural Biology 1994; 4:868-874.
19. Bohacek, R.S., McMartin, C. and Guida, W.C.: The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. Medicinal Research Reviews 1996; 16:3-50.
20. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanochi, T. and Tasumi, M.: The Protein Data Bank: a computer based archival file for macromolecular structures. J. Mol. Biol. 1977; 112:532-542.
21. Oliver, S.G., *et al.*: The complete DNA sequence of yeast chromosome III. Nature 1992; 357:38-46.
22. Anfinsen, C.B.: Principles that govern the folding of protein chains. Science 1973; 181:223-230.
23. Lesk, A.M. In: Kent A, Williams GJ, Hall CM andKent R (eds.) Encyclopedia of computer science and technology 31, Supplemnt 16. New York, Basel, Hong Kong: Marcel Dekker, Inc., 1994;
24. Moult, J., Pedersen, J.T., Judson, R. and Fidelis, K.: Results of the 1994 Structure Prediction Competition and meeting 'Critical assessment of techniques for protein structure prediction'. Proteins 1995; 23:ii-iv.
25. Chothia, C. and Lesk, A.M.: The relation between divergence of sequence and structure in proteins. EMBO J. 1986; 5:823-826.
26. Mosimann, S., Meleshko, R. and James, M.N.G.: A critical assessment of comparative molecular modeling of tertiary structures of proteins. PROTEINS: Structure, Function and Genetics 1995; 23:301-317.
27. Pizzi, E., Tramontano, A., Tomei, L., La Monica, N., Failla, C., Sardana, M., Wood, T. and De Francesco, R.: Molecular model of the specificity pocket of the hepatitis C virus protease: Implications for substrate recognition. Proc. Natl. Acad. Sci. 1994; 91:888-892.
28. Benson, D.A., Boguski, M.S., Lipman, D.J. and Ostell, J.: Genbank. Nucl. Acids. Res. 1997; 25:1-6.
29. Stoesser, G., Sterk, P., Tuli, M.A., Stoehr, P.J. and Cameron, G.N.: The EMBL nucleotide sequence database. Nucl. Acids. Res. 1997; 25:7-13.
30. Tateno, Y. and Gojobori, T.: DNA Data Bank of Japan in the age of information biology. Nucl. Acids. Res. 1997; 25:14-17.
31. George, D.G., Dodson, R.J., Garavelli, J.S., Haft, D.H., Hunt, L.T., Marzec, C.R., Orcutt, B.C., Sidman, K.E., Srinivasarao, G.Y., Yeh, L.S.L., Arminski, L.M., Ledley, R.S., Tsugita, A. and Barker, W.C.: The Protein Information Resource (PIR) and the PIR-international protein sequence database. Nucl. Acids. Res. 1997; 25:24-27.
32. Bairoch, A. and Apweiler, R.: The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucl. Acids. Res. 1997; 25:31-36.
33. Allen, F.H., Bergerhoff, G. and Sievers, R. Crystallographic Databases. in International Union of Crystallography. Chester, UK: 1987.
34. Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Smith, J.M. and Watson, D.G.: The development of version 3 and 4 of the Cambridge Structural Database System. J. Chem. Inf. Comput. Sci. 1991; 31:187-204.

35. Bairoch, A., P., B. and Hofmann, K.: The PROSITE database, its status in 1995. Nucl. Acids Res. 1996; 24:189-196.
36. Henikoff, S. and Henikoff, J.G.: Automated assembly of protein blocks for database searching. Nucl. Acids. Res. 1991; 19:6565-6572.
37. Kabsch, W. and Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983; 22:2577-2637.
38. Sander, C. and Schneider, R.: Database of homology-derived protein structures and the structural meaning of sequence alignment. PROTEINS: Structure, Function and Genetics 1991; 9:56-68.
39. Holm, L., Ouzounis, C., Sander, C., Tuparev, G. and Vriend, G.: A database of protein structure families with common folding motifs. Protein Science 1992; 1:1691-1698.
40. Needleman, S.B. and Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 1970; 48:443-453.
41. Sankoff, D. and Kruskal, J.B. Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Reading, MA: Addison-Wesley, 1983.
42. Devereux, J., Haeberli, P. and Smithies, O.: A comprehensive set of sequence analysis programs for the VAX. Nucl. Acids Res. 1984; 12:387-395.
43. Dayhoff, M.O., Schwartz, R.M., Chen, H.R., Barker, W.C., Hunt, L.T. and Orcutt, B.C. In: Dayhoff MO (eds.) Atlas of Protein Sequence and Structure 5 (3). Washington D.C.: National Biomedical Research Foundation, 1981;
44. Vingron, M.: Multiple Sequence Alignments and Applications in Molecular Biology. 1991, Naturwissenschaftlich-mathematischen Gesamtfakultat der Ruprecht-Karls-Universitat: Heidelberg.
45. Henikoff, S. and Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 1992; 89:10915-10919.
46. Pearson, W.R. and Lipman, D.J.: Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 1988; 85:2444-2448.
47. Altschul, S.F. and Lipman, D.J.: Protein database searches for multiple alignments. Proc. Natl. Acad. Sci. USA 1990; 87:5509-5513.
48. Pearson, W.R.: Comparison of methods for searching protein sequence databases. Protein Science 1995; 4:1145-1160.
49. Moult, J.: The current state of art in protein structure prediction. Current Opinion in Biotechnology 1996; 7:422-427.
50. Tramontano, A. In: Villar HO (eds.) Advances in Computational Biology 2. Greenwich: JAI Press, 1995; 239-259.
51. Chung, S.Y. and Subbiah, S. In: Hunter L and Klein T (eds.) First Pacific Symposium on Biocomputing, Cona, Haway Singapore: World Scientific, 1996; 126-141.
52. Venkatachalam, C.: Stereochemical criteria for polypeptides and proteins. V. Conformation of three linked peptide units. Biopolymers 1968; 6:1425-1436.
53. Rose, G.D., Gierasch, L.M. and Smith, J.A.: Turns in peptides and proteins. Advances in Protein Chemistry 1985; 37:1-109.
54. Sibanda, B.L. and Thornton, J.M.: ß-hairpin families in globular proteins. Nature 1985; 316:170-174.
55. Efimov, A.V.: Standard conformation of polypeptide chain in irregular regions of proteins. Mol. Biol. (USSR) 1986; 20:250-260.

56. Leszczynski, J.F. and Rose, G.D.: Loops in globular proteins: a novel category of secondary structure. Science 1986; 234:849-855.
57. Milner-White, E.J., Ross, B.M., Ismail, R., Belhadi-Mostefa, K. and Poet, R.: One type of gamma-turn rather than the other gives raise to chain-reversal in proteins. J. Mol. Biol. 1988; 204:772-782.
58. Wilmot, C. and Thornton, M.J.: Analysis and prediction of the different types of beta turns in proteins. J. Mol. Biol. 1988; 203:221-232.
59. Sbanda, B.L., Blundell, T.L. and Thornton, J.M.: Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. J. Mol. Biol. 1989; 206:759-777.
60. Creighton, T.E. Proteins: structures and molecular properties. San Francisco: Freeman, W.E. and Co., 1993.
61. Chothia, C. and Lesk, A.M.: Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol. 1987; 196:901-917.
62. Chothia, C., Lesk, A.M., Levitt, M., Amit, A.G., Mariuzza, R.A., Philips, S.E.V. and Poljak, R.: The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. Science 1987; 196:901-918.
63. Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W. and Colman, P.: The conformation of immunoglobulin hypervariable regions. Nature 1989; 342:877-883.
64. Tramontano, A., Chothia, C. and Lesk, A.M.: Framework residue 71 is a major determinant of the second hypervariable region in VH domains of immunoglobulins. J. Mol. Biol. 1990; 215:175-182.
65. Lesk, A.M. and Tramontano, A. Antibody structure and structural predictions useful in guiding antibody engineering. New York: Freeman & Co., 1990.
66. Searle, S.J., Pedersen, J.T., Henry, A.H., Webster, D.M. and Rees, A.R. In: Borreback CAK (eds.) Antibody Engineering Oxford: Oxford University Press, 1994; 3-51.
67. Shirai, H., Kidera, A. and Nakamura, H.: Structural classification of CDR-H3 in antibodies. FEBS Letters 1996; 399:1-8.
68. Morea, V., Tramontano, A., Rustici, M., Chothia, C. and Lesk, A.M.: Antibody structure, prediction and redesign. Biophysical Chemistry (in press);
69. Tramontano, A. and Lesk, A.M.: Common features of the conformations of antigen-binding loops in immunoglobulins and applications to modelling of loop conformations by data base screening. Proteins 1992; 13:231-245.
70. Karplus, M. and Weaver, D.L.: Protein-folding dynamics. Nature 1976; 260:404-406.
71. Lesk, A.M. and Rose, G.D.: Folding units in globular proteins. Proc. Natl. Acad. Sci. U. S. A. 1981; 78:4304-4308.
72. Martin, A.C.R., Cheetham, J. and Rees, A.R.: Modeling antibody hypervariable loops: A combined algorithm. Proc. Natl. Acad. Sci. U. S. A. 1989; 86:9268-9272.
73. Montelione, G.T. and Scheraga, H.A.: Formation of local structures in protein folding. Acc. Chem. Res. 1989; 22:70-76.
74. Skolnick, J. and Kolinski, A.: Computer simulations of globular protein folding and tertiary structure. Ann. Rev. Phys. Chem. 1989; 40:207-235.
75. Moult, J. and Unger, R.: An analysis of protein folding pathways. Biochemistry 1991; 30:3816-3824.

76. Tramontano, A., Chothia, C. and Lesk, A.M.: Structural determinants of the conformations of medium sized loops in proteins. Proteins 1989; 6:382-394.
77. Kabsch, W. and Sander, C.: On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. U. S. A. 1984; 81:1075-1978.
78. Benner, S.A.: Predicting de novo the folded structure of proteins. Curr. Opin. Struct. Biol. 1992; 2:402-412.
79. Cohen, B.I. and Cohen, F.E. Prediction of protein secondary and tertiary structure. New York: Academic Press, 1994.
80. Janin, J., Wodak, S., Levitt, M. and Maigret, B.: Conformation of amino acid side chains in proteins. J. Mol. Biol. 1978; 125:357-386.
81. Dunbrack, R.L. and Karplus, M.: Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nature Structural Biology 1994; 1:334-339.
82. Vasquez, M.: Modeling side-chain conformation. Current Opinion in Structural Biology 1996; 6:217-221.
83. Peitsch, M.C.: Protein modelling by e-mail. Bio/Technology 1995; 13:658-660.
84. Sali, A., Potterton, L., Yuan, F., van Vlijmen, H. and Karplus, M.: Evaluation of comparative protein modeling by MODELLER. Proteins: Structure Function and Genetics 1995; 23:318-326.
85. Connolly, M.L.: Analytical molecular surface calculations. J. Appl. Crystallography 1983; 16:548-558.
86. Richards, F.M.: Calculation of molecular volumes and areas for structures of known geometry. Methods Enzymol. 1985; 115:440-464.
87. Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M.: PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst. 1993; 26:283-291.
88. Murzin, A.G., Brenner, S.E., Hubbard, T.J.P. and Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 1995; 247:536-540.
89. Chothia, C.: One thousands families for the molecular biologist. Nature 1993; 357:543-544.
90. Lemer, C.M.-R., Rooman, M.J. and Wodak, S.J.: Protein Structure Prediction by Threading Methods: Evaluation of Current Techniques. Proteins 1995; 23:337-355.
91. Jones, D.T.: Progress in protein structure prediction. Current Opinion in Structural Biology 1997; 7:377-387.
92. Finkelstein, A.V.: Protein structure: what is it possible to predict now? Current Opinion in Structural Biology 1997; 7:60-71.
93. Bowie, J.U., Luthy, R. and Eisenberg, D.: A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991; 253:164-170.
94. Bowie, J.U. and Eisenberg, D.: Inverted protein structure prediction. Curr. Opin. Struct. Biol. 1993; 3:437-444.
95. Luthy, R., Bowie, J.U. and Eisenberg, D.: Nature 1992; 356:83-85.
96. Godzik, A., Kolinski, A. and Skolnick, J.: Topology fingerprint approach to the inverse protein folding problem. J. Mol. Biol. 1992; 227:227-238.
97. Jones, D.T., Taylor, W.R. and Thornton, J.M.: A new approach to protein fold recognition. Nature 1992; 358:86-89.

98. Sippl, M.J. and Weitckus, S.: Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. PROTEINS: Structure, Function and Genetics 1992; 13:258-271.
99. Lathrop, R.H.: The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Engineering 1994; 7:1059-1068.
100. Bryant, S.H. and Altschul, S.F.: Statistics of sequence-structure threading. Current Opinion in Structural Biology 1995; 5:236-244.
101. Sippl, M.J.: Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. Journal of Computer-Aided Molecular Design 1993; 7:473-501.
102. Jones, D.T. and Thornton, J.M.: Potential energy functions for threading. Current Opinion in Structural Biology 1996; 6:210-216.
103. Jernigan, R.L. and Bahar, I.: Structure-derived potentials and protein simulations. Current Opinion in Structural Biology 1996; 6:195-209.
104. Finkelstein, A.V., Badretdinov, A.Y. and Gutin, A.M.: Why do protein architectures have Boltzmann-like statistics? PROTEINS: Structure, Function and Genetics 1995; 23:142-150.
105. Thomas, P.D. and Dill, K.A.: Statistical potentials extracted from protein structure - how accurate are they? J. Mol. Biol. 1996; 257:457-479.
106. Sippl, M.J.: Knowledge-based potentials for proteins. Current Opinion in Structural Biology 1995; 5:229-235.
107. Hubbard, T.J. and Park, J.: Fold recognition and ab initio structure predictions using Hidden Markov Models and ?-strand pair potentials. PROTEINS: Structure, Function and Genetics 1995; 23:398-402.
108. Rost, B. In: Rawlings CJ, Clark D, Altman R, Lengauer T andWodak S (eds.) Proc. of the Third International Conference on Intelligent Systems for Molecular Biology Menlo Park: AAAI Press, 1995; 314-321.
109. Russell, R.B., Copley, R.R. and Barton, G.J.: Protein fold recognition by mapping predicted secondary structures. J. Mol. Biol. 1996; 259:349-365.
110. Rost, B., Schneider, R. and Sander, C.: Progress in protein structure prediction? Trends Biochem. Sci. 1993; 18:120-123.
111. Eisenberg, D.: Into the black of night. Nature Structural Biology 1997; 4:95-97.
112. Defay, T. and Cohen, F.E.: Evaluation of current methods for ab initio protein structure prediction. Proteins 1995; 23:431-445.
113. Nishikawa, K., Kubota, Y. and Ooi, T.: Classification of proteins into groups based on amino acid composition and other characters. I Angular distribution. J. Biochem. 1983; 94:981-995.
114. Sheridan, R.P., Dixon, J.S. and Venkataraghavan, R.: Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. Biopolymer 1985; 24:1995-2023.
115. Muskal, S.M. and Kim, S.-H.: Predicting protein secondary structure content. J. Mol. Biol. 1992; 225:713-727.
116. Gibson, T.J., Postma, J.P., Brown, R.S. and Argos, P.: A model for the tertiary structure of
the 28 residue DNA-binding motif
('zinc finger') common to many

eukaryotic transcriptional regulatory
proteins. Prot. Eng. 1988; 2:209.
117. Barton, G.J.: Protein secondary structure prediction. Current Opinion in Structural Biology 1995; 5:372-376.
118. Rost, B. and Sander, C.: Progress of 1D protein structure prediction at last. PROTEINS: Structure, Function and Genetics 1995; 23:295-300.
119. Zvelebil, M.J., Barton, G.J., Taylor, W.R. and Sternberg, M.J.E.: Prediction of protein secondary structures and active sites using alignments of homologous sequen ces. J. Mol. Biol. 1987; 195:957-961.
120. Levin, J.M., Pascarella, S., Argos, P. and Garnier, J.: Quantification of secondary structure prediction improvement using multiple alignments. Prot. Eng. 1993; 6:849-854.
121. Salamov, A.A. and Solovyev, V.V.: Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J. Mol. Biol. 1995; 247:11-15.
122. Janin, J.: Surface and inside volumes in globular proteins. Nature 1979; 277:491-492.
123. Hubbard, T.J. and Blundell, T.L.: Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modeling. Protein Eng. 1987; 1:159-171.
124. Miller, S., Janin, J., Lesk, A.M. and Chothia, C.: Interior and surface of monomeric proteins. J. Mol. Biol. 1987; 196:641-656.
125. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. and Zehfus, M.H.: Hydrophobicity of am ino acid residues in globular proteins. Science 1985; 229:834-838.
126. Sander, C., Scharf, M. and Schneider, R. In: Rees AR, Sternberg MJE andWetzel R (eds.) Protein Engineering Oxford: IRL Press, 1992; 89-115.
127. Rost, B. and Sander, C.: Conservation and prediction of solvent accessibility in protein families. Proteins: Structure, Function and Genetics 1994; 20:216-226.
128. Eddy, S.R.: Hidden Markov models. Curr. Opin. Struct. Biol. 1996; 6:361-365.
129. Benner, S.A., Gerloff, D. and Chelvanayagam, G.: The phospho-?-galactosidase and synaptotagmin predictions. Proteins 1995; 23:446-453.
130. Rost, B. and Sander, C.: Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 1993; 232:584-599.
131. Hubbard, T.J. In: Lathrop RH (eds.) Proceedings of the Biotechnology Computing Track, Protein Structure Prediction MiniTrack of the 27th HICSS V. IEEE Computer Society Press, 1994; 336-354.
132. Lifson, S. and Sander, C.: Specific recognition in the tertiary structure of ?-sheets of proteins. J. Mol. Biol. 1980; 139:627-639.
133. Gobel, U., Sander, C., Schneider, R. and Valencia, A.: Correlated mutations and residue contacts in proteins. PROTEINS: Structure, Function and Genetics 1994; 18:309-317.
134. Thomas, D.J., Casari, G. and Sander, C.: The prediction of protein contacts from multiple sequence alignments. Protein Engineering 1996; 9:941-948.
135. Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G.H. and Benner, S.A.: An analysis of simultaneous variation in protein structures. Protein Engineering 1997; 10:307-316.
136. Lesk, A.M. and Chothia, C.: The response of protein structures to amino acid sequence changes. Phil. Trans. Roy. Soc. London 1986; 317:345.
137. Casari, G., Sander, C. and Valencia, A.: A method to predict functional residues in proteins. Nature Structural Biology 1995; 2:171-178.

138. Pedersen, J.T. and Moult, J.: Genetic algorithms for protein structure prediction. Current Opinion in Structural Biology 1996; 6:227-231.
139. Pedersen, J.T. and Moult, J.: Ab initio protein structure prediction for small polypeptides and protein fragments using genetic algorithms. Proteins: Structure, Function and Genetics 1995; 23:454-460.
140. Finkelstein, A.V., Gutin, A.M. and Badretdinov, A.Y.: Perfect temperature for protein structure prediction and folding. Proteins: Structure, Function and Genetics 1995; 23:151-162.
141. Hagler, A.T. and Honig, B.: On the formation of protein tertiary structure on a computer. Proc. Natl. Acad. Sci. U. S. A. 1978; 75:554-558.
142. Levitt, M.: Protein folding by restrained energy minimization and molecular dynamics. J. Mol. Biol. 1983; 170:723-764.
143. Moult, J. and James, M.N.: An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1986; 1:146-163.
144. Moser, R., Frey, S., Münger, K., Hehlgans, T., Klauser, S., Langen, H., Winnacker, E.-L., Mertz, R. and B., G.: Expression of the synthetic gene of an artificial DDT-binding polypeptide in *Escherichia coli*. Protein Eng. 1987; 1:339-343.
145. Mütter, M.: Nature's rules and chemist's tools: a way for creating novel proteins. *Trends Biochem. Sci.* 1988; 13:260-265.
146. Goraj, K., Renard, A. and Martial, J.A.: Synthesis, purification and initial structural characterization of octarellin, a *de novo* polypeptide modelled on the alpha/beta-barrel proteins. Protein Eng. 1990; 3:259-266.
147. Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. and Hecht, M.H.: Protein design by binary patterning of polar and non polar amino acids. Science 1993; 262: 1680-1685.
148. Hill, C.P., Anderson, D.H., Wesson, L., DeGrado, W.F. and Eisenberg, D.: Crystal structure of ? -1: implications for protein design. Science 1990; 249:543-546.
149. Raleigh, D.P. and DeGrado, W.F.: A *de novo* designed protein shows a thermally induced transition from a native to a molten globul-like state. *J. Am. Chem. Soc.* 1992; 114:10079-10081.
150. Lovejoy, B., Seunghyon, C., Cascio, D., McRorie, D.K., DeGrado, W.F. and D., E.: Crystal structure of a synthetic triple-stranded ? -helical bundle. Science 1993; 259:1288-1293.
151. Fersht, A. and Winter, G.: Protein Eng. *Trends Biochem. Sci.* 1992; *17*:292-294.
152. Jones, D.T.: De novo protein design using pairwise potentials and a genetic algorithm. Prot. Sci. 1994; 3:567-574.
153. Hubbard, T., Park, J., Lahm, A., Leplae, R. and Tramontano, A.: Protein structure prediction: playing the fold. Trends In Biochemical Sciences 1996; 21:279-281.
154. Hubbard, T. and Tramontano, A.: Update on protein structure prediction: results of the 1995 IRBM workshop. Folding and Design 1996; 1:R55-63.
155. Arevalo, J.H., Stura, E.A., Taussig, M.J. and Wilson, I.A.: Three-dimensional structure of an anti-steroid Fab' and progesterone-Fab' complex. J. Mol. Biol. 1993; 231:103-118.
156. Satow, Y., G.H. Cohen, Padlan, E.A. and Davies, D.R.: Phosphocholine binding immunoglobulin Fab McPC603: an X-ray diffraction study at 2.7 Å J. Mol. Biol. 1986; 190.:593-604.

Figure legends

Fig. 1: Modeling by homology of the variable domain of antibodies: superposition of the model structure of antibody DB3 to the crystal structure [155]. The model and the structure are cyan and green in the framework region, yellow and orange in the H3 region red and magenta in the other hypervariable loops, respectively. Only the C, C? and N atoms of the main chain are shown.

Fig. 2: Designed structure of the Minibody: the design was based on the structure of the variable domain of an antibody of known structure. The 61-residue minibody protein includes three ?-strands from each of the two ?-sheets of the variable (V) heavy chain domain of the mouse antibody McPC603 [156], along with the segments corresponding to the exposed hypervariable H1 and H2 loops of the immunoglobulin as defined by Chothia and Lesk [61]. A metal binding site, shown in the Figure, was designed in the molecule in order to probe its proper folding.

Fig. 3: a) Window of the fold recognition analysis tool in GLASS. The output of a fold recognition program is reformatted and analysed. Each fold on the data base, shown with its respective score can be selected to create a starting model. b) The model can be displayed using RasMol (Sayle, R.: Rasmol. 1994; See <URL: `http://www.glaxowelcome.cs.uk/netscape/software/`>. The ribbon represents the variability of the sequence calculated from a multiple sequence alignment: the more variable residues have a larger red ribbon, the more conserved have a small blue ribbon. Dotted lines represent predicted correlated mutations. GLASS allows to map experimental data on a predicted structure by user defined coloured balls and/or labels. c) View of the multiple sequence alignment used to generate the ribbon in (b), coloured according the hydrophobicity of the residues.

Predicting Protein Disorder for N-, C- and Internal Regions

Xiaohong Li[1] xli@eecs.wsu.edu
Pedro Romero [1] promero@eecs.wsu.edu
Meeta Rani [2] meeta_wsu@hotmail.com
A. Keith Dunker [2] dunker@disorder.chem.wsu.edu
Zoran Obradovic [1] zoran@eecs.wsu.edu

[1] School of Electrical Engineering and Computer Science
Washington State University, Pullman, WA 99164, U.S.A.
[2] School of Molecular Biosciences
Washington State University, Pullman, WA 99164, U.S.A.

Abstract

Logistic regression (LR), discriminant analysis (DA), and neural networks (NN) were used to predict ordered and disordered regions in proteins. Training data were from a set of non-redundant X-ray crystal structures, with the data being partitioned into N-terminal, C-terminal and internal (I) regions. The DA and LR methods gave almost identical 5-cross validation accuracies that averaged to the following values: 75.9 ± 3.1% (N-regions), 70.7 ± 1.5% (I-regions), and 74.6 ± 4.4% (C-regions). NN predictions gave slightly higher scores: 78.8 ± 1.2% (N-regions), 72.5 ± 1.2% (I-regions), and 75.3 ± 3.3% (C-regions). Predictions improved with length of the disordered regions. Averaged over the three methods, values ranged from 52% to 78% for length = 9-14 to ≥ 21, respectively, for I-regions, from 72% to 81% for length = 5 to 12-15, respectively, for N-regions, and from 70% to 80% for length = 5 to 12-15, respectively, for C-regions. These data support the hypothesis that disorder is encoded by the amino acid sequence.

1 Introduction

The current paradigm is that protein function depends on 3D structure [10, 16, 18], yet some proteins are partially or completely unfolded in their native states [2, 3, 7, 24, 26]. For such "natively unfolded" [30], "natively disordered" [9] or "intrinsically unstructured" [31] proteins, the lack of a fixed 3D structure can be an integral part of the function. Are such disordered proteins common or rare?

To estimate the commonness of disordered proteins, we applied predictors of disorder to appropriate databases [20]. The results suggested that intrinsic disorder is common [21], but lack of structural information limits confidence in these findings. Since the needed structural information will be slow in coming, we are revisiting the question of commonness by improving our disorder predictions.

A limitation of our previous studies was that only neural networks (NNs) were tried. By comparing NNs with discriminant analysis (DA) and logistic regression (LR), we can gain additional confidence in the suitability of prediction for identifying ordered and disordered protein.

Technical limitations of our previous algorithms resulted in absence of predictions on 15 residues at each end [20], resulting in non-prediction of a significant fraction of the residues. Here we modified the algorithms to extend the predictions to the N- and C-termini.

2 Materials and Methods

2.1 Data

Using missing electron density in X-ray structures as indicating disorder [19], we identified 115 N-terminal, 84 C-terminal and 69 internal (I) disordered regions (DRs) that were contained in 197

unrelated proteins listed in PDB-select-25 [11]. The minimum lengths used were 5 and 9 for termini and I-regions, respectively. The various DRs contained the following numbers of residues 1,644 (N-regions), 1,347 (I-regions) and 1,250 (C-regions). A set of 130 unrelated, disorder-free proteins that were also from PDB-select-25 [11] was used to generate the ordered residues used for predictor training.

2.2 Attribute Generation

Composition-based and property-based attributes were calculated over sliding windows [20, 32] . A total 51 attributes were examined, where the sets of amino acids represented some property such as aromaticity, charge, sheet formers, etc (Table 1).

Table 1: Attributes list.

Var.	Attributes	Var.	Attributes	Var.	Attributes	Var.	Attributes
X1	FWY	X14	WCFIYVLHM	X27	WY	X40	P
X2	FWY(H/2)	X15	ATRGQSNPDEK	X28	A	X41	Q
X3	KR-D-E	X16	WYFAS	X29	C	X42	R
X4	KR-D-E(H/2)	X17	WYFKR	X30	D	X43	S
X5	KRDE	X18	WYFKRH	X31	E	X44	T
X6	KRDE(H/2)	X19	WYFDE	X32	F	X45	V
X7	WFYC	X20	WYFEDH	X33	G	X46	W
X8	WFYC(H/2)	X21	FWYKRDE	X34	H	X47	Y
X9	STQHNDERK	X22	FWYKRDEH	X35	I	X48	PEVK
X10	WEYCVILMP	X23	EMAL	X36	K	X49	Flexibility
X11	VILM	X24	YNPG	X37	L	X50	Hydropathy
X12	STQHN	X25	VIYFW	X38	M	X51	Coordination number
X13	GSA	X26	SGKPDE	X39	N		

Composition-based attributes were the sums of the numbers of the indicated amino acids in a given window. For example, aromaticity, X1 = FWY, the number of phenylalanines (F) + tryptophans (W) + tyrosines (Y) within a given window. The number of histidines was sometimes divided by 2 (e.g. H/2) due to its small ring size or partial charge. For the net charge attributes, X3 and X4, the number of each negative residue was subtracted (e.g. -D, -E) from the number of positive residues.

Property-based attributes were the sums the residue property-values. For X49 = flexibility, the value for each residue was based on its backbone-atom B-factors averaged over 92 unrelated protein structures [28]. The values for X50 =l hydropathy were from the Kyte-Doolittle scale [15]. X51 = coordination number is the average number of side chain neighbors that are in contact with the given side chain when it is fully buried as determined from a set of 33 non-homologous proteins [8].

As in previous studies [20], a window of 21 was used for I-regions. A window of 11 was used for positions 6 onwards and for -6 backwards for N- and C-regions, respectively. Predictions at positions 1 to 5 and -1 to -5 used windows of size 6 to 10, respectively. For N-regions, these windows included residues from the end to 5 positions beyond the position being predicted, and for C-regions, from the end to 5 positions before.

2.3 Logistic Regression Model and Attribute Selection

The logistic regression (LR) model was developed for dealing with the situation in which the dependent variable is binary [5]. Here we used order = 0, and disorder = 1. SAS (Release 6.12, SAS Institute,

Cary, NC) was used for the calculations.

For a given threshold probability, an observation is classified into the category with the probability higher than the threshold. In the logistic model, the probability is estimated from the following equation:

$$ln[\frac{p}{1-p}] = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_j X_{ij}$$

where $p = P$ ($Y_i = 1$ for ordered) and $1 - p = P$ ($Y_i = 0$ for disordered); $i = 1, 2, \cdots, n$, where n is the sample size; $j = 1, 2, \ldots, m$, where m is the attribute number; and $X_{i1}, \ldots, X_{ij}$ are attributes used for prediction.

The parameters b_i are estimated by maximizing the following function:

$$\sum_{i=1}^{n} P(B, Y_i) = \sum_{i=1}^{n} ln(\frac{1}{1 - e^{-BX_i}})$$

where B is the vector of parameters need to be estimated. After all b_i values are estimated, p can be calculated as:

$$p = \frac{1}{1 + e^{-BX_i}}$$

For order = 0 and disorder = 1, the threshold is set to be 0.5; if $p \geq 0.5$, then the amino acid is predicted to be disordered; otherwise, ordered. The LR is applied each time an attribute is introduced or removed, and the Chi-square test is executed [1]. The process is repeated until introduction or removal of an attribute leads to no change at a significance level of 0.05. Eight selected attributes were used in LR predictor even though a few more number passed the significance test.

2.4 Discriminant Analysis Model

For discriminant analysis (DA), it is assumed that prior probabilities are equal, that the variables (attributes) are independent, and that all attribute values satisfy the normal distribution. Since we used sliding windows to obtain data and since many of the attributes share dependencies on the same amino acids, the assumption that the data are independent is not true. However, this lack of independence didn't seem to cause serious problems since this approach gave results comparable to the other methods in this study. Again, SAS (Release 6.12, SAS Institute, Cary, NC) was used to carry out the calculations for this model.

For the ordered and disordered data $\chi = \{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, n$; $y_i = \{0, 1\}$, where $y = 0$ for an ordered amino acid, $y = 1$ for a disordered one. The x_i values are the attributes data. We used Bayesian discriminant analysis method to predict the probability that a given amino acid belongs to an ordered or disordered regions. The posterior (conditional) probability that a residue belongs to an ordered or disordered region is given by the following equation:

$$P(C_j \mid x) = \frac{P(x \mid D)P(D)}{P(x \mid D)P(D) + P(x \mid O)P(O)}$$

where $j = 0$ (ordered) or 1 (disordered); $P(O)$ and $P(D)$ are the a priori probabilities of a residue being ordered and disordered residues, respectively. $P(x \mid D)$ and $P(x \mid O)$ are the conditional densities of disordered and ordered residues, respectively. $P(C_j \mid x)$ is given by the following relationship:

$$P(C_j \mid x) = \frac{e^{C_{j0} + \mathbf{b}_j' \mathbf{x}}}{\sum_{k=1}^{m} e^{(b_{k0} + \mathbf{b}_k' \mathbf{x})}} = \frac{1}{1 + e^{(b_{d0} - b_{o0}) + (\mathbf{b_d} + \mathbf{b_o})' \mathbf{x}}}$$

Using observed data, the parameters b_{d0} and b_{o0} and the vectors $\mathbf{b_d}$ and $\mathbf{b_o}$ can be estimated. The classification for a given pattern x is determined as: $Class = \arg\max\{P(C_j \mid x)\}$, where class is 0 or 1 for ordered or disordered, respectively.

The attributes were repeatedly introduced or removed, and the F-test was applied after each operation, until no attributes could be introduced or removed at a significance level of 0.05 [6]. The top eight selected attributes were used for establishing the DA predictor even though a greater number were accepted at the significance level indicated.

2.5 Neural Network Model

The application of NNs to order/disorder prediction has been described elsewhere in more detail [20]. The feed forward NN used in this study is fully connected with an 8x8x1 architecture, which has eight inputs (selected by LR), one hidden layer with 8 nodes and one output layer with one node. The back propagation method was used for data training [23].

3 Results

3.1 Attribute selection

A list of 51 attributes was used in this study (Table 1). Many of the attribute values are correlated. In addition, some attributes make little contribution in distinguishing the ordered and disordered regions. Finally, 51 attributes is simply too many for the amount of disordered data. These characteristics necessitated the selection of a subset of the attributes for the predictors.

Stepwise DA and stepwise LR were used for attribute selection on ordered and disordered data from the N-, C- and I- regions. Although more than 8 attributes were selected for the data at a significance level of 0.05, the ninth and later selected attributes make relatively little contribution, as shown by the prediction accuracy upon addition of attributes in their order of importance (Fig. 1).

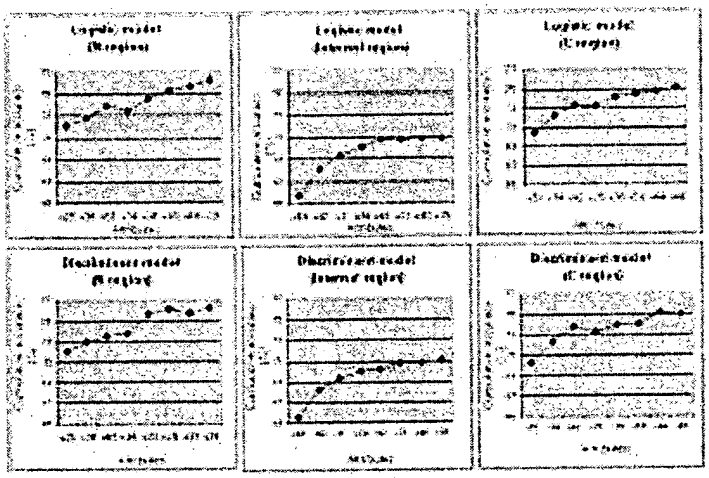

Figure 1: Contribution of selected attributes on prediciton.

The selected attributes in Table 2 start with the most important. For the top 8 sequence attributes in a given protein region, the DA and LR models selected almost the same ones. That is, 5/8, 6/8, and 8/8 attributes were selected in common by the two methods for the N-, I-, and C-region data, respectively. In contrast, the selected attributes were very different for the different regions. Only 1 sequence attribute was selected in common for all three regions. For the 3 pairs of regions, only 4/8 were selected in common for N- and C-regions, just 3/8 for the N- and I-regions, and a mere 2/8 for the C- and I-regions. These results suggest that sequence characteristics leading to disorder depend on the location of the region in the sequence.

Table 2: Attributes selected according to the significance in DA and LR.

Attributes	1	2	3	4	5	6	7	8
DA: N-terminal region	X25	X38	X51	X34	X20	X35	X31	X39
LR: N- terminal region	X25	X38	X51	X34	X30	X45	X48	X39
DA: Internal region	X49	X42	X11	X34	X43	X31	X40	X35
LR: Internal region	X49	X42	X7	X14	X43	X31	X40	X35
DA: C-terminal region	X51	X34	X42	X25	X38	X50	X44	X48
LR: C-terminal region	X51	X34	X42	X25	X38	X50	X44	X48

3.2 Prediction Accuracies

The prediction accuracies of the 3 models over the 3 regions are given in Table 3. The DA and LR models gave almost identical accuracies for each region, with the largest difference being 0.3% (for I-regions). Also, using the N-regions as an example, the 0.1% difference between the two methods is much less than the ± 3.5% and ± 2.7% variation among the 5-cross validation trials. Thus, the DA and LR models give essentially indistinguishable prediction accuracies overall.

Table 3: Five-cross validations of the predictors developed by three methods.

Model	Region	1	2	3	4	5	Average
Neural Network	N region	79.0%	78.8%	78.7%	78.9%	78.7%	78.8% (±1.2%)
	I region	72.2%	72.6%	73.1%	72.2%	72.4%	72.5% (±1.2%)
	C region	75.1%	75.5%	74.9%	74.4%	76.5%	75.3% (±3.3%)
Discriminant Analysis	N region	74.2%	78.4%	75.9%	73.7%	77.2%	75.9% (±3.5%)
	I region	70.1%	71.3%	70.0%	71.8%	71.1%	70.9% (±1.4%)
	C region	72.7%	71.6%	77.0%	76.3%	75.9%	74.7% (±4.1%)
Logistic Regression	N region	74.0%	77.3%	76.3%	74.2%	77.2%	75.8% (±2.7%)
	I region	69.6%	70.62%	69.8%	71.7%	71.4%	70.6% (±1.6%)
	C region	72.0%	71.3%	77.3%	76.6%	75.5%	74.5% (±4.7%)

The NN approach gives slightly higher predictions for all three regions. In the following, the first number in each pair is the NN accuracy and the second number is the average of the DA and LR accuracies: 78.8 ± 1.2% versus 75.9 ± 3.1% (N-regions), 72.5 ± 1.2% versus 70.7 ± 1.5% (I-regions), and 75.3 ± 3.3% versus 74.6 ± 4.4% (C-regions).

3.3 Cross Prediction

Each predictor was applied to the data from the regions not used for its training, here called cross prediction. In Table 4 accuracies observed during 5-cross validation (indicated by *) are compared with the accuracies for cross predictions (no *). For the most part, as expected, the accuracy of a given predictor drops when applied to the data from a region different from its training set. However, for both the LR and DA models trained on I-regions, the accuracies remain essentially the same when the predictors are applied to C-region data. That is, the LR model only changes from 70.6% on its I-regions training data to 70.9% when applied to C-region data, and the DA model, from 70.9% to 71.2%. This failure to drop in accuracy is especially surprising since I- and C-regions predictors share just 2/8 attributes.

Table 4: Cross-prediction specificity for disordered regions.

Predictors	Region	N-terminal Data	Internal DR Data	C-terminal Data
Discriminant Model	N-region	75.9%*	52.9%	61.5%
	Internal region	64.9%	70.9%*	71.2%
	C-region	71.3%	68.8%	74.7%*
Logistic Model	N-region	75.8%*	44.6%	57.6%
	Internal region	66.3%	70.6%*	70.9%
	C-region	71.6%	68.9%	74.5%*

3.4 Length dependence of prediction accuracy.

To estimate accuracy versus length, the prediction outputs were partitioned according to length with the number of residues in each class indicated in parenthesis (Table 5). For the DA and LR predictions in Table 5, the models from 5-cross validation were retrained on 5/5 of the data, whereas for the NN predictions, retraining on the whole set of data was not performed. Instead, one of the NN models, which was trained on 4/5 of the data, was used. For DRs of 9 to 14, the roughly 52% accuracy (averaged over the 3 methods) corresponds to essentially random classification. For DRs of 15 to 20, the average accuracy increased to 74%, and for DRs $\geq$ 21 the average increased still further to about 78%. Since the windows are 21 in length, the shorter DRs fill only a fraction of their windows, and therefore the poor accuracies are expected.

Table 5: Prediction accuracies for different I-DR lengths.

Predictors	9-14 AA (379)	15-20 AA (262)	21AA or longer (707)
NN	52.8%	73.7%	78.6%
DA	50.9%	74.4%	77.9%
LR	52.2%	74.4%	78.2%

The lowered prediction rates due to the short disordered windows probably helps to explain the surprising cross prediction results that occur when the predictors trained on I-regions are applied to C-region data as described above.

The N- and C-region data also show length-dependent accuracies (Table 6). For N-region data, the accuracies, averaged over the three methods, change from 72% (length = 5), to 83% (length = 6-8), to 77% (length = 9-11) to 81% (length = 12-15). For C-region data, the respective averaged accuracies are 69%, 78%, 72% and 80%.

Table 6: Prediction accuracies for different N- or C-DR lengths.

DR Regions	Predictors for N and C regions	DR=5 AA (N:60; C:65)	DR=6-8AA (N:269; C:117)	DR=9-11 AA (N219; C135)	DR=12-15AA (N:137; C:163)
N	NN	75.0%	83.6%	77.1%	86.0%
	DA	71.7%	83.3%	78.1%	81.0%
	LR	70.0%	82.2%	76.3%	77.4%
C	NN	70.5%	73.1%	74.2%	85.2%
	DA	67.7%	74.4%	63.0%	75.5%
	LR	67.7%	74.4%	63.0%	76.1%

3.5 Position-by-position accuracy for N- and C-regions

The position-by-position error rates were determined; all three predictors give similar outputs that result in fairly complex curves (Fig. 2). The data in Fig. 2 are incommensurate with the data in Table 6, so these should not be compared directly. This is discussed below in more detail.

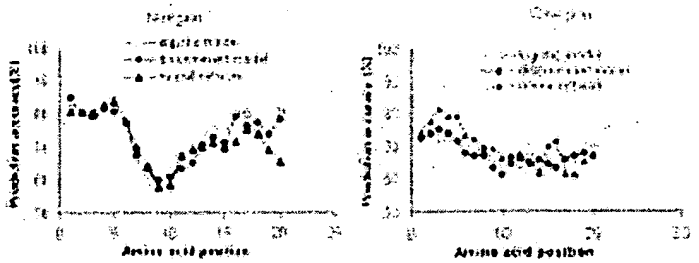

Figure 2: Prediction accuracy over AA positions in N- and C- regions.

4 Discussion

4.1 Data

Disorder characterized by X-ray diffraction can be static or dynamic [13]. In our previous studies we attempted to remove this ambiguity by finding independent information such as protease sensitivity or NMR spectra [20], but most often the information was lacking. As an alternative, we compared X-ray-characterized disorder with disorder characterized by other methods especially NMR [9]. The results indicate that ambiguity of X-ray characterized disorder is not fatal, but probably leads to the introduction of noise into the training data.

4.2 Comparison of Prediction models

There is no single best algorithm for pattern recognition problems. Performance for a given algorithm depends on the data set being investigated [14]. DA, LR and NN approaches are among the most commonly used, and all have been applied to sequence analysis problems. DA has been successfully used for predicting internal exons of DNA sequences [25] and protein secondary structure [27, 33]. LR has been used for identifying regulatory regions of genes [29]. NNs have also been used for predicting secondary structure [22]. Considering the characteristics of the three methods, we decided to try all of them in this study.

The LR and DA models exhibited nearly identical performance for the disorder predictions whereas the NN gave a slightly higher accuracy (Table 3). Application of Cochran's test [4] indicates a real significance for the superiority of the neural network. However, prediction accuracy is a simplistic indicator, so it seems inappropriate to rank the methods on this basis alone.

Olson [17] reported that, with proper selection of attributes, both statistical and neural network classifiers yield essentially identical accuracies for a given test case. From this, there are two implications that arise from the possible superiority of the NN predictors. First, other factors not included in Table 1 might affect the determination of order or disorder. To test this, other attributes need to be investigated. Alternatively, the predictors might not be optimized.

DA is fast and performs well except for very skewed data [14]. LR was developed for binary data and so might be the most robust for predicting two states, order and disorder. DA and LR methods need much less computation time than NN, and produce results that are easier to interpret.

Back propagation NN, in most cases, performs well especially for noisy data. Noisy data is of particular concern due to the ambiguity of X-ray-characterized disorder. With appropriate architecture,

a back propagation neural network can be a universal approximator for arbitrary finite inputs [12]. No assumptions are required for the input and output parameters.

There are some general disadvantages, however, in using NNs. For example, the selection of the architecture (number of layers, number of neurons) is empirical. If too few hidden neurons are used, training convergence is often poor, whereas if too many are used, the network might converge well, but generalization is typically poor. A further shortcoming of NNs is the failure to provide insight. That is, there is no deterministic way to carry out attribute selection. For these reasons, we carried out an entirely separate study to gain understanding of our problem [32]. A significant advantage of the LR and DA methods is the ability to carry out step-wise addition of the various attributes.

4.3 Attribute Selection

Both our previous studies and the studies on I-region data presented here used windows of 21 residues. Despite the very different databases in the two studies, the previously selected attributes closely resemble those reported here. That is, 6 of 8 attributes were selected in common by the LR and DA methods; these were X49 (flexibility), X42 (R), X43 (S), X31 (E), X40 (P), and X35 (I) as shown in Table 2. Of the 6 attributes in common, 5 were selected in our previous studies on completely different databases of order and disorder; only the last, and least important attribute found here, X35, was not selected previously. Of the 4 attributes not selected in common, e.g. X11 (VILM) and X34 (H) by DA and X7 (WFYC) and X14 (WCFIYVLHM) by LR, all are identical to, or share amino acids with, attributes selected previously on completely different data.

The prediction of order or disorder for I-regions depends on a balance of different types of attributes. X49 (flexibility), X42 (R), X43 (S), X31 (E), and X40 (P) are attributes that, at high value, favor disorder, whereas X35 (I), X11 (VILM), X7 (WFYC), and X14 (WCFIYVLHM) all favor order.

This is the first study of the relationship between amino acid sequence and disorder at the ends of proteins. Comparing attributes for N- and C-regions with each other and with attributes for I-regions provides insight regarding disorder at the ends of proteins.

Although just 4/8 attributes are in common between the two ends, these include the top two attributes for each (Table 2). That is, the top two attributes, X25 (VIYFW) and X 38 (M), for N-regions data rank fourth and fifth, respectively, for C-regions data. Also, the first, X51 (Coordination Number), and second, X34 (H) for C-regions rank third and fourth, respectively, for N-regions. From Fig. 1, these top attributes are the most important. Of the attributes specific for each end, some of these contain residues with charges opposite to the charge at the termini (Table 2). For example, the positive charge at the N-terminus is opposite to the negative charges (E and D) in X20 (WYFEDH) and to that of X31 (E). Likewise, the negative charge of the C-terminus is opposite to the positive charge of X42 (R).

The attributes selected for the N- and C-regions can for the most part be described as being associated with the formation of ordered structure, whereas the attributes selected for I-regions appear to be more balanced between attributes favoring order and those favoring disorder. Even the charged attributes, X31 (E), and X42 (R), which are associated with disorder in I-regions, are selected at the ends in a manner that brings about charge balance and so could be promoting order in these regions. Perhaps I-regions are neutral with respect to order or disorder, whereas perhaps N- and C-regions tend to be naturally disordered. If so, order or disorder in I-regions is determined by the overall balance of various types of attributes, whereas overcoming the natural disorder tendency at the ends may require the presence of order-inducing amino acids in these regions.

4.4 Prediction accuracies

If only the longer I-regions data are considered, the estimated accuracy here (Table 5) is slightly better than we found earlier. That is, here we find about 78% (average of DA and LR) versus about 73% - 74% (NN) reported previously [20]. The slight improvement probably relates to the increased number

of attributes surveyed, 51 here versus 24 previously. More specifically, only single amino acids were used in the original study, whereas the expanded set used here contains combinations of amino acids. Several of the selected combinations include groups of the single amino acids selected in the original study, thus creating space for additional inputs that bring more information to bear on the problem.

The length-dependence of I-region predictions shows a very large gradient, from almost random predictions (near 52% averaged over the three methods) for length = 9-14 to fairly strong predictions (about 78% averaged over the three methods) for length $\geq$ 21. Because windows of 21 were used, the shorter lengths only partially filled the windows and so the essentially random predictions are a reasonable outcome when the disorder training examples contain large amounts of order.

Here we report our first attempt to predict to the ends of the protein. We included down to very short DRs (5 amino acids) with the expectation that we would find some minimum length below which the predictions would fail completely. Such failure would give random predictions like those observed for the shortest I-regions data, although for different reasons. To our surprise, even DRs as short as 5 amino acids at the ends yielded good prediction accuracies, about 72% (N-regions) and 70% (C-regions) when averaged over the three methods (Table 6). Although not monotonic, increases in accuracy reached 82% (N-regions) and 80% (C-regions) for DRs of length 12-15. These high values suggest the possibility of special effects at the ends of proteins.

The NN, LR, and DA methods give similar curves for the position-dependent accuracies at each end (Fig. 2), with high value followed by minima that are very noticeable for the N-region curves and barely noticeable for the C-regions curves. The causes of these minima near positions 9-10 are uncertain. One possibility is that windows at the 9-10 positions for the disorder data contain substantial fractions of ordered residues, resulting from a combination of the distribution of disorder lengths in the training data and the way in which the windows were specified. Based on this idea, we are exploring alternative window specifications with the goal of reducing these minima.

The data in Fig. 2 were grouped differently from the data in Table 2. This leads to false discrepancies such as the > 80% accuracies for positions 1-5 (N-regions, Fig. 2) which appear to be better than the 72% accuracy for N-region DRs = 5 AA (averaged over the 3 methods from Table 6). The false discrepancy arises because the data for Table 6 come from the specified lengths whereas the data for Fig. 2 are predictions at particular positions from DRs of all different lengths. So, the higher accuracy of > 80% for the first 5 positions results from contributions from DRs longer than 5, which yield predictions over the first 5 positions better than the 72% observed for DRs of length = 5.

4.5 Implications for Future Research

The high accuracy of prediction of very short DRs at the termini might be special, due to end effects, or the high accuracy might be simply the result of the use of very short windows. If the latter is true, then use of shorter windows might be of benefit for I-region predictions as well.

A second task will be to merge our various predictors into one, making it possible to predict disorder from the amino to the carboxyl terminus of a protein. This will open the way for a variety of projects, such as improving predictions of disorder on a genomic basis and such as using disorder predictions to indicate which proteins are likely to crystallize and which ones are not.

Acknowledgments

Support from NSF research grant NSF-CSE-IIS-9711532 to Z. Obradovic and A. K. Dunker is gratefully acknowledged. Dr. R. Drossu's neural network simulator was used in part of the study.

References

[1] Anderson, E.B., *The Statistical Analysis of Categorical Data, 2nd ed.*, Springer-Verlag, 1991.

[2] Bloomer, A.C., Champness, J.N., Bricogne, G., Staden, R., and Klug, A., Protein disk of tobacco mosaic virus at 2.8Åresolution showing the interactions within and between subunits, *Nature*, 276:362–368, 1978.

[3] Bode, W., Schwager, P., and Huber, R., The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9Åresolution, *J. Mol. Biol.*, 118(1):99–112, 1978.

[4] Cochran, W.G., The comparison of percentages in matched samples, *Biometrika*, 37:256–266, 1950.

[5] Cox, D.R., *The Analysis of Binary Data*, London, Methuen and Co., 1970.

[6] Eisenbeis, R.A. and Avery, R.B., *Discriminant Analysis and Classification Procedures: Theory and Applications*, Lexington, Mass. Health., 1972.

[7] Fletcher, C.M. and Wagner, G., The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein, *Protein Sci.*, 7(7):1639–1642, 1998.

[8] Galaktionov, S.G. and Marshall, G.R., Prediction of Protein Structure in Terms of Intraglobular Contacts: 1D to 2D to 3D, Technical Report, Center for Molecular Design, Washington University, St. Louis, 1996.

[9] Garner, E., Cannon, P., Romero, P., Obradovic, Z., and Dunker, A., Predicting disordered regions from amino acid sequence: common theme despite differing structural characterization, *Genome Informatics*, 9:201–214, 1998.

[10] Hagerman, P.J.I., From sequence to structure to function, *Curr. Opin. Struct. Biol.*, 6(3):277–280, 1996.

[11] Hobohm, U. and Sander, C., Enlarged representative set of protein structures, *Protein Sci.*, 3(3):522–524, 1994.

[12] Hornik, K., Approximation capabilities of multilayer feedforward networks, *Neural network*, 4:251–257, 1991.

[13] Huber, R., Conformational flexibility in protein molecules, *Nature*, 280:538–539, 1979.

[14] King, R.D., Feng, C., and Sutherland, A., Statlog: comparison of classification algorithms on large real-world problems, *Applied Artificial Intelligence*, 9:289–333, 1995.

[15] Kyte, J. and Doolittle, R.F., A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, 157(1):105–132, 1982.

[16] Mirsky, A.E. and Pauling, L., On the structure of native, denatured and coagulated proteins, *Proc. Natl. Acad. Sci. USA*, 22:439–447, 1936.

[17] Olson, K.M. and Ybarra, G.A., A performance comparison of neural network and statistical pattern recognition approaches to automatic target recognition of ground vehicles using SAR imagery, Miceli, W.J., *Proceeding of SPIE: Radar Processing, Technology, and Applications II*, San Diego, CA, USA, August 1997, 3161:159–170, 1997.

[18] Orengo, C.A. and Todd, A.E., From protein structure to function, *Curr. Opin. Struct. Biol.*, 9:374–382, 1999.

[19] Rani, M.P.R., Obradovic, Z., and Dunker, A.K., Annotation of PDB with respect to "disordered regions" in proteins, *Genome Informatics*, 9:240–241, 1998.

[20] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., and Dunker, A.K., Identifying disordered regions in proteins from amino acid sequences, *Proc. I.E.E.E. International Conference on Neural Networks,* 1:90–95, 1997.

[21] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Guilliot, S., Garner, E., and Dunker, A.K., Thousands of proteins likely to have long disordered regions, *Pacific Symposium on Biocomputing*, 3:435–446, 1998.

[22] Rost, B. and Sander, C., Improved prediction of protein secondary structure by use of sequence profiles and neural networks, *Proc. Natl. Acad. Sci. USA*, 90(16):7558–7562, 1993.

[23] Rumelhart, D.E., Durbin, R., Dolden, R., and Chauvin, Y., Backpropagation: the basic theory, Y. Chauvin and D.E. Rumelhart (eds): *Back Propagation: Theory and Applications*, Hillside, NJ: Lawrence Erlbaum:1–34, 1985.

[24] Schulz, G.E., Nucleotide Binding Proteins, Molecular Mechanism of Biological Recognition, Elsevier/North-Holland Biomedical Press, 79–94, 1979.

[25] Solovyev, V.V., Salamov, A.A., and Lawrence, C.B., Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames, *Nucleic Acids Res.*, 22(24):5156–5163, 1994.

[26] Spolar, R.S. and Record II, M.T., Coupling of local folding to site-specific binding of proteins to DNA, *Science*, 263:777–784, 1994.

[27] Stolorz, P., Lapedes, A., and Xia, Y., Predicting protein secondary structure using neural net and statistical methods, *J. Mol. Biol.*, 225(2):363–377, 1992.

[28] Vihinen, M., Torkkila, E., and Riikonen, P., Accuracy of protein flexibility predictions, *PROTEINS: Struct. Funct. Genet,* 19(2):141–149, 1994.

[29] Wasserman, R., Felix, C.A., McKenzie, S.E., Shane, S., Lange, B., and Finger, L.R., Identification of an altered immunoglobulin heavy-chain gene rearrangement in the central nervous system in B-precursor acute lymphoblastic leukemia, *Leukemia*, 7(8):1294–1299, 1993.

[30] Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A., and Lansbury, P.T., Jr., NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded, *Biochemistry*, 35(43):13709–13715, 1996.

[31] Wright, P.E. and Dyson, H.J., Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm, *J. Mol. Biol.*, 293(2):321–331, 1999.

[32] Xie, Q., Arnold, G.E., Romero, P., Obradovic, Z., Garner, E., and Dunker, A.K., The sequence attribute method for determining relationships between sequence and protein disorder, *Genome Informatics*, 9:193–200, 1998.

[33] Zhang, X., Mesirov, J.P., and Waltz, D.L., Hybrid system for protein secondary structure prediction [published erratum appears in J Mol Biol 1993 Aug 20;232(4):1227], *J. Mol. Biol.*, 225(4):1049–1063, 1992.